Logic, Epistemology, and the Unity of Science 28

Fabrice Pataut *Editor*

# Truth, Objects, Infinity

## New Perspectives on the Philosophy of Paul Benacerraf

Springer

# Logic, Epistemology, and the Unity of Science

Volume 28

*Logic*, *Epistemology*, *and the Unity of Science* aims to reconsider the question of the unity of science in light of recent developments in logic. At present, no single logical, semantical or methodological framework dominates the philosophy of science. However, the editors of this series believe that formal techniques like, for example, independence friendly logic, dialogical logics, multimodal logics, game theoretic semantics and linear logics, have the potential to cast new light on basic issues in the discussion of the unity of science.

This series provides a venue where philosophers and logicians can apply specific technical insights to fundamental philosophical problems. While the series is open to a wide variety of perspectives, including the study and analysis of argumentation and the critical discussion of the relationship between logic and the philosophy of science, the aim is to provide an integrated picture of the scientific enterprise in all its diversity.

More information about this series at http://www.springer.com/series/6936

Fabrice Pataut
Editor

# Truth, Objects, Infinity

## New Perspectives on the Philosophy of Paul Benacerraf

Springer

*Editor*
Fabrice Pataut
FRE Sciences, Normes, Décision, CNRS
Paris
France

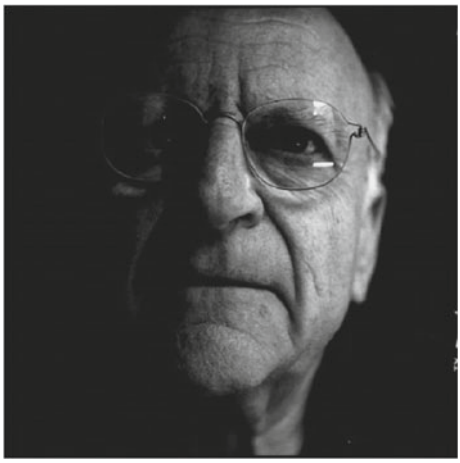
ISSN 2214-9775 ISSN 2214-9783 (electronic)
Logic, Epistemology, and the Unity of Science
ISBN 978-3-319-45978-3 ISBN 978-3-319-45980-6 (eBook)
DOI 10.1007/978-3-319-45980-6

Library of Congress Control Number: 2016950889



Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland



# Acknowledgements

In many ways, the articles "What Numbers Could Not Be" and "Mathematical Truth" by Paul Benacerraf have dominated the philosophy of mathematics for about fifty years. If you take into account the early version of "Mathematical Truth" that Benacerraf has kindly agreed to publish in this volume for the first time, some of his main ideas have circulated in the philosophical community since the late sixties.

Many seminal articles and important books have argued in favor of views about what truth in mathematics amounts to and about what numbers could or could not be. Most of them have done so by taking a stand on these two timeless papers; at the very least, this is where they started from. The philosophy of mathematics might be philosophy "in a pure state, stripped of all worldly appendages; austere; philosophy without the sugar coating of a pretense to Relevance to Life" (Benacerraf 1996b[1]: 10), yet everyone working today in philosophy *tout court* has read them.

And this is only, if I may say so, the tip of the iceberg. "Frege: The Last Logicist," from 1981, and Benacerraf's very first publication, "Tasks, Super-Tasks, and Modern Eleatics," from 1962, have also shaped, or reshaped, some key areas of the discipline. Because of them, people have looked at Frege's logicist project of reducing mathematics to logic, and at the vexed puzzle of how an infinite number of operations might be performed in a finite time, in a different way.

As far as Frege is concerned, we should perhaps travel back to the spring of 1961, when Benacerraf presented his 1960 dissertation on the "fundamentally mistaken" doctrine of logicism in a course on the philosophy of mathematics he gave at Princeton (Benacerraf 1960: iii, 255). George Boolos, then a student, was "irritated" by Benacerraf's reading of it, judged it "perverse" but later on came to agree with it (Boolos 1996: 143–144)[2]. It is striking that Benacerraf's rejection of logicism, mainly on the ground that concepts of mathematics cannot be defined in terms of concepts of pure logic (Chaps. I and II of Benacerraf 1960), and that we do not accept (or reject) mathematics on extra-mathematical grounds (Chap. III of Benacerraf 1960), has stimulated an interest in the logicist outlook, e.g., in sub-programs of logicism offering reductions of fragments of arithmetic, in neo-logicism and, more generally, in a discussion of Hume's principle.

As for supertasks, Benacerraf and Putnam pointed as early as 1964, in the wake of the 1962 paper, that "[i]f we take the stand that 'nonconstructive' procedures—i.e., procedures that require us to perform infinitely many operations in a finite time—are conceivable,[3] though not *physically* possible (owing mainly to the existence of a limit to the velocity with which physical operations can be performed), then we can say that there does 'in principle' exist a verification/refutation procedure for number theory" (Benacerraf and Putnam [1964] 1983: 20). The point here is that the notion of truth in number theory is a bona fide notion only insofar as the notion of a completed actually infinite series of operations is one itself. Intuitionist worries about the actual infinite, Dummettian antirealist worries about the non-constructivity of classical proofs, and, a fortiori, strict finitist worries about the permissiveness of the "in principle" clause at work in the claim, all take us back to the notion of truth, a unifying, perhaps *the* unifying theme of Benacerraf's work.[4]

Then, last but not least, there is the anthology of texts in the philosophy of mathematics, the *Selected readings* that generations of students have used as a sourcebook—indeed as a textbook—edited, twice prefaced, and introduced in collaboration with Hilary Putnam (Benacerraf and Putnam [1964] 1983).

Few philosophers trigger such a fresh start on perennial philosophical problems and do so much to teach their discipline. The idea of an international meeting devoted to a critical evaluation of Benacerraf's work in the philosophy of mathematics, with its wide bearings on the philosophy of language, the philosophy of logic, and epistemology, seemed a natural one. The idea came in the form of a workshop rather than in the form of a glorifying celebration, with a large part of the time to be set aside for discussion. The prospect of benefiting from Paul's participation certainly gave it more than a tinge of excitement.

The workshop eventually took place in Paris at the Collège de France on May 10 and May 11, 2012. The participants were Jody Azzouni, Jacques Dubucs, Bob Hale, Brice Halimi, Sébastien Gandon, Mary Leng, Andrea Sereni, Stewart Shapiro, Claudine Tiercelin, and myself. The present volume includes some of the papers read at the workshop and five additional essays. Unfortunately, the papers by Jacques Dubucs, Bob Hale and Claudine Tiercelin are missing from the volume. Hale's "Properties and the Interpretation of Second-Order Logic" has since appeared in *Philosophia Mathematica*, Vol. 21 (2), 2013, pp. 133–156; it was first published online by the journal on August 3, 2012. The five essays that were not presented at the workshop are, respectively, by Justin Clarke-Doane, Jon Pérez Laraudogoitia, Antonio León-Sánchez and Ana C. León-Mejía, Marco Panza, and Philippe de Rouilhan.

The fourteen essays have been conveniently organized into four parts. Mine serves as an introduction. The first part includes those that directly address what is known today as "Benacerraf's dilemma." The second gathers contributions directly involved with issues in the philosophy of mathematics and in particular with that part of it concerned with arithmetic and number theory. The third is devoted to supertasks. The fourth part contains the ancestor of "Mathematical Truth," a draft from January 1968 that Benacerraf started writing in 1967 with the joint support of the John Simon Guggenheim Foundation and of Princeton University, and

"Comments on Reduction," a lecture he gave in a graduate seminar at Princeton around 1975 to discuss the views on reductions of natural numbers within set theory that he previously expounded in print in "What Numbers Could Not Be." Many thanks to Paul for having allowed the publication of these two essays.

I would like to thank the Institut d'Histoire et de Philosophie des Sciences et des Techniques for its administrative and financial support. Special thanks are due to its director, Jean Gayon, who enthusiastically backed up the project from the very beginning; to Marco Panza, for his scientific and financial help through the Université Panthéon-Sorbonne (Paris 1) and the Agence Nationale de la Recherche; and to Peggy Cardon Tessier, Alexandre Roulois, and Sandrine Souraya Lucigny for their precious help and patience in all administrative and organizational matters. I would also like to thank Claudine Tiercelin who holds the Chaire de métaphysique et de philosophie de la connaissance at the Collège de France, for participating and for being a most gracious host during these two days. Many thanks indeed to Jean-Marie Chevalier and Benoît Gaultier, respectively maître de conférences and attaché temporaire d'enseignement de recherche at the Collège de France at the time of the workshop, for making the participants' and the audience's life so enjoyable on this occasion. The Agence Nationale de la Recherche, the Centre National de la Recherche Scientifique, the Université Panthéon-Sorbonne (Paris 1) and the Università Vita e Salute San Raffaele, Milan, must also be thanked for their generous financial support.

The Paul Benacerraf Workshop could not have taken place without them.

Last but not least, I would like to thank the series editors, Shahid Rahman and John Symons, who have welcomed the project of a volume devoted to the work of Paul Benacerraf, and to Christi Lue and Hema Suresh, the editorial assistants for history, philosophy of science and logic at Springer, for their patience and faultless expertise in all editorial matters.

October 2016 Fabrice Pataut

## Notes

1. Benacerraf's name, either by itself or with the name of another author, followed by a date, refers to the corresponding publication in the Chronological Bibliography of Paul Benacerraf to 2016 at the end of this volume.
2. Boolos, G. (1996). On the Proof of Frege's Theorem. In A. Morton and S. P. Stich (Eds.), *Benacerraf and his Critics* (pp. 143–159). Blackwell Publishers: Oxford and Cambridge, Mass.
3. E.g., if one has an infinite series of operations to perform, say $S_1$, $S_2$, $S_3$, … and if one is able to perform $S_1$ in 1 min, $S_2$ in 1/2 min, $S_3$ in 1/4 min, etc.; then in 2 min one will have completed the whole infinite series.
   [The note appears as footnote 12 of the Introduction [revised] in Benacerraf and Putnam [1964] 1983: 20. *Editor's note*].
4. See, e.g., "Mathematical Truth (1968 version)," in this volume: Sect. 12.1, and the *Answer to objection 2: Truth vs. knowledge* in Sect. 12.5.

# Contents

# Editor and Contributors

## About the Editor

**Fabrice Pataut** is a researcher at the CNRS (FRE Sciences, Normes, Décision). His publications include "Naturalizing Mathematics and Naturalizing Ethics" (*Ontological Landscapes*, V. Petrov, ed., Ontos Verlag, 2011, pp. 183–210), "Réalisme, sens commun et langage ordinaire" (*John L. Austin et la philosophie du langage ordinaire*, S. Laugier et C. Al-Saleh, éds., Georg Olms Verlag, 2011, pp. 417–432) and "Comments on 'Parsimony and Inference to the Best Mathematical Explanation' (Reply to Baker)" (*Synthese*, Vol. 193(2), 2016, pp. 351–363). He is the guest editor, with Daniele Molinini and Andrea Sereni, of the special issue of *Synthese* on *Indispensability and Explanation* (*Synthese*, Vol. 193 (2), 2016). He specializes in the philosophy of language, the philosophy of logic, and the philosophy of mathematics.

## Contributors

**Jody Azzouni** is currently at Tufts University. He works in metaphysics, philosophy of mathematics, philosophy of logic, philosophy of language, and philosophy of science. His publications include *Talking about Nothing: Numbers, Hallucinations, and Fictions* (Oxford UP, 2010) and *Semantic Perception: How the Illusion of a Common Language Arises and Persists* (Oxford UP, 2013).

**Paul Benacerraf** taught at Princeton University since he joined the faculty in 1960. He held many key administrative positions there and was appointed Stuart Professor of Philosophy in 1979 and James S. McDonnell Distinguished University Professor of Philosophy in 1998. He was elected a fellow of the American Academy of Arts and Sciences in 1988. "What Numbers Could Not Be" (*The Philosophical Review*, Vol. 74, 1965, pp. 47–73) and "Mathematical Truth" (*The Journal of Philosophy*, Vol. 70, 1973, pp. 661–679) are among the most influential

articles of twentieth-century philosophy. He is the editor, with Hilary Putnam, of *Philosophy of mathematics: Selected readings* (2nd ed., Cambridge UP, 1983).

**Justin Clarke-Doane** is currently an assistant professor at Columbia University, an Honorary Research Fellow at the University of Birmingham, and a Honorary Research Associate at Monash University in Australia. He graduated from New York University with a Ph.D. in philosophy on *Morality and Mathematics*. He is the author of "Multiple Reductions Revisited" (*Philosophia Mathematica*, Vol. 16, 2008, pp. 244–255) and "Morality and Mathematics: The Evolutionary Challenge" (*Ethics*, Vol. 122, 2012, pp. 313–340). He is the editor of the special issue of the *Review of Symbolic Logic* on objectivity and undecidability in mathematics (*RSL*, Vol. 5, 2012).

**Sébastien Gandon** is currently a professor of Philosophy at the Université Blaise Pascal in Clermont-Ferrand. He is the director of the research center Philosophie et Rationalités and the author of *Russell's Unknown Logicism* (Palgrave, 2012). He specializes in the history of analytic philosophy, the history of mathematics, and the philosophy of mathematics.

**Brice Halimi** currently teaches at the Université de Paris Ouest, where he is a member of the Institut de Recherches Philosophiques. His recent publications include "The Versatility of Universality in *Principia Mathematica*" (*History and Philosophy of Logic*, Vol. 32, 2011, pp. 241–264) and "Diagrams as Sketches" (*Synthese*, Vol. 186, 2012, pp. 387–409). He is the author of *Le nécessaire et l'universel — Analyse et critique de leur corrélation* (Librairie philosophique Jean Vrin, 2013).

**Mary Leng** is a lecturer of Philosophy at the University of York. She is the author of *Mathematics and Reality* (Oxford UP, 2010), which offers a defence of fictionalism in the philosophy of mathematics.

**Ana C. León-Mejía** did her graduate work in neuroscience and the biology of human behavior and has a Ph.D. in sociology. She is a lecturer of Developmental Psychology at the International University of La Rioja, Spain, and an associate lecturer in Psychology at the Open University in London. She is the editor of a *Handbook of Extracurricular Activities for Students with Special Needs* (CEP Publishers, 2014) and of *Empathy Deficit Disorder in School Contexts* (San Pablo CUE Press, 2014).

**Antonio León-Sánchez** is an independent researcher. He has completed graduate studies in geology and postgraduate studies in mathematics, logic, and philosophy of science. His papers on the actual infinity hypothesis, supertasks, ω-order, and the special theory of relativity have been gathered in two volumes: *Infinity at Stake—Selected Arguments on the Actual Infinity Hypothesis* (Bubok Publishing, 2013) and *Digital Relativity—A Digital Reinterpretation of the Special Theory of Relativity* (Bubock Publishing, 2014).

**Marco Panza** is a senior researcher at the CNRS (Institut d'Histoire et de Philosophie des Sciences et des Techniques). He is the author of several books and articles in the history of mathematics, especially early modern and eighteenth-century mathematics, and in the philosophy of mathematics, especially on platonism. These include *Newton et les origines de l'analyse : 1664–1666* (Blanchard, 2005) and, with Andrea Sereni, *Plato's Problem: An Introduction to Mathematical Platonism* (Palgrave MacMillan, 2013).

**Jon Pérez Laraudogoitia** is an associate professor at the University of the Basque Country (UPV/EHU) in the Department of Logic and Philosophy of Science. His publications include the "Supertask" entry of the *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/entries/spacetime-supertasks, 1999; substantive revision, 2009), "The New String Paradox" (*Philosophy of Science*, Vol. 80, 2013, pp. 143–154) and "The Supertask Argument Against Countable Additivity" (*Philosophical Studies*, Vol. 168 (3), 2014, pp. 619–628).

**Philippe de Rouilhan** is a senior researcher emeritus at the CNRS and a member of the Institut d'Histoire et de Philosophie des Sciences et des Techniques of which he was the director for a long time. His work pertains to logic *lato sensu* or, more specifically, to formal ontology, formal semantics, the philosophy of mathematics and the philosophy of language. He is currently working on truth and logical consequence, and on hyperintensional logic. His numerous publications in either French or English include *Russell et le cercle des paradoxes* (PUF, 1996)—reviewed in *Journal of Symbolic* Logic, Vol 64, 1999, pp. 1845–1847, by N. C. A. da Costa and O. Bueno—,"Russell's Logics" (*Lecture Notes in Logic*, Vol. 19, 2005, pp. 335–349) and, with Paul Gochet, *Logique épistémique et philosophie des mathématiques* (Dunod, 2007).

**Andrea Sereni** is currently an associate professor at the Scuola Universitaria Superiore Pavia. He has taught philosophy at the universities of Bologna and Modena and at the Università Vita e Salute San Raffaele in Milan. His publications include, in collaboration with Marco Panza, *Plato's Problem: An Introduction to Mathematical Platonism* (Palgrave MacMillan, 2013), "Indispensability Arguments and their Quinean Heritage" (*Disputatio*, Vol. 32, 2012, pp. 343–360) in collaboration with Jacob Busch, and "Frege, Indispensability and the Compatibilist Heresy," *Philosophia Mathematica*, Vol. 23 (1), 2015, pp. 11–30. His areas of expertise are the philosophy of mathematics and epistemology.

**Stewart Shapiro** is currently the O'Donnell Professor of Philosophy at The Ohio State University. He specializes in logic, philosophy of logic, philosophy of language, and philosophy of mathematics. His publications include *Vagueness in context* (Oxford UP, 2006, reissued in paperback, 2008), "The company kept by cut abstraction (and its relatives)" (*Philosophia Mathematica*, Vol. 19, 2011, pp. 107–138), and "An 'i' for an i: singular terms, uniqueness, and reference" (*Review of Symbolic Logic*, Vol. 5, 2012, pp. 380–415).

# Introduction
# In Defence of a Princess Margaret Premise

## 1. Introductory Remarks

In their introduction to the volume *Benacerraf and his Critics*, Adam Morton and Stephen Stich remark that "[t]wo bits of methodology will stand out clearly in anyone who has talked philosophy with Paul Benacerraf": (i) "[i]n philosophy you never prove anything; you just show its price," and (ii) "[f]ormal arguments yield philosophical conclusions only with the help of hidden philosophical premises" (Morton and Stich 1996: 5).

The first bit will come to some as a disappointment and to others as a welcome display of suitable modesty. Frustration notwithstanding, the second bit suggests that, should one stick to modesty, one might after all prove something just in case one discloses the hidden premises of one's chosen argument and pleads convincingly in their favor on independent grounds.

I would like to argue that a philosophical premise may be uncovered—of the kind that Benacerraf has dubbed "Princess Margaret Premise" (PMP)[1]—that helps us reach a philosophical conclusion to be drawn from an argument having a metamathematical result as one of its other premises, viz., Gödel's first incompleteness theorem. To be somewhat more (im)precise, something philosophical may be inferred from Gödel's THEOREM VI (Gödel [1931] 1986: [187] 173) (and its proof) with respect to the vexed question whether truth may transcend recognizability in principle by us, humans.

Obviously, this first approximation of the question in such unblushingly Dummettian terms of transcendence, truth, recognizability, and the in principle *vs.* effective distinction must be found wanting. It is, of course, unspecific to a fault to speak in such general terms of *the* grand philosophical problem of the relation between truth and the recognition of truth, but it is not for that matter either dreadfully vague or offensively inexact. The argument I shall propose, provided it is

indeed one, has three virtues: (i) It shows with a non-negligible degree of accuracy what is the price one has to pay for the philosophical conclusion I believe may be secured; (ii) it is genuinely philosophical (as opposed to genuinely metamathematical) and does indeed lead to a conclusion, so that something in philosophy may be proved after all; (iii) thanks to (i)–(ii), the argument enables us to come up with a more refined version of the problem I just started with without thereby loosing sight of its very general scope. I shall explain why it is important not to loose sight of such a scope in the concluding remarks, i.e., why, although powerful formal tools play a crucial role in the obtaining of a sober and strictly philosophical conclusion, the larger philosophical picture, for the benefit of which "a lot of delicate informal interpretation" (Morton and Stich loc. cit.) has to be put to good use, *must* in the end matter to us (see infra Sect. 8).

The only rather feeble apology I am able to offer at this point with regard to the unreliability of the introductory formulation of the question to be answered is that the forthcoming argument reveals "unexpected premises and consequences" (Morton and Stich loc. cit.) at play in the controversy about the independence of truth from our ability to recognize that truth obtains when it does, so that, if the argument does indeed go through, we shall at least end up with an improved formulation of the philosophical puzzle we started with.[2]

## 2. Correctness and Truth

Let me start with Gödel's result proper and with remarks—some by Gödel, some by others—pertaining to it that will play a role in the forthcoming argument.

Gödel certainly thought it worthwhile to remind his readers that his proof of the first incompleteness theorem was *constructive*. He pointed out the fact that the result had been obtained "in an intuitionistically unobjectionable manner" (Gödel [1931] 1986: [189] 177) and offered as a warrant for this claim that "all existential statements [*Existentialbehauptungen*] occurring in the proof [were] based upon Theorem V [i.e., the theorem immediately preceding the first incompleteness theorem] which, as is easily seen, is unobjectionable from the intuitionistic point of view" (Gödel loc. cit.: note 45a). In Kleene's terminology, Theorem V states that every primitive recursive relation is numeralwise expressible in *P*, where *P* is the system obtained from Whitehead and Russell's *Principia Mathematica*, without the ramification of the types, taking the natural numbers as the lowest type and adding their usual Peano axioms (Kleene 1986: 129). When expressed formally, without reference to any particular interpretation of the formulas of *P*, and in Gödel's own terminology which favors the indirect talk of Gödel numbers and of concepts

applying to these numbers rather than a direct talk of the formal objects (i.e., the formulas and the variables), THEOREM V claims that:

> For every recursive relation $R(x_1,\ldots, x_n)$ there exists an $n$-place RELATION SIGN r (with the FREE VARIABLES $u_1$, $u_2,\ldots, u_n$) such that for all $n$-tuples of numbers $(x_1,\ldots, x_n)$ we have
>
> $$R(x_1,\ldots, x_n) \rightarrow \text{Bew}\,[Sb(\mathrm{r}^{u_1 \ \ldots \ u_n}_{Z(x_1)\ldots\ Z(x_n)})],$$
>
> $$\overline{R}(x_1,\ldots, x_n) \rightarrow \text{Bew}\,[\text{Neg}(Sb(\mathrm{r}^{u_1 \ \ldots \ u_n}_{Z(x_1)\ldots\ Z(x_n)}))].$$
>
> Gödel [1931] (1986: [186] 171)

Gödel sketches an outline of the proof and notes on this occasion that THEOREM V is itself "of course, […] a consequence of the fact that in the case of a recursive relation $R$ it can, for every $n$-tuple of numbers, be decided *on the basis of the axioms of the system P* whether the relation $R$ obtains or not" (Gödel op. cit.: [186n39] 171n39). This, it must be noted, may also be decided by means of procedures that remain unobjectionable from the intuitionistic standpoint.

Gödel's true and undecidable formula, the existence of which is proved constructively by the first incompleteness theorem (THEOREM VI), may seem at first sight to offer a counterexample to the claim that truth may not transcend recognition by us, either in principle or effectively, for the proof establishes the existence of a formula which does have both properties, viz., that of truth *and* that of undecidability. Gödel's diagonal argument does indeed provide a true statement which is nevertheless omitted by the relevant algorithm.[3]

Two aspects of the situation passed on to us by Gödel's proof somewhat complicate the matter. First, there are followers of Wittgenstein's *Remarks on the Foundations of Mathematics* like Shanker who think that it is incoherent and indeed downright nonsensical to claim that a statement or formula is (or may be, for that matter) both true and undecidable. Shanker points out that, if a complete manifestation of our recognition of the truth of the Gödel formula were possible, the semantic formulation of the theorem would thereby be defective: It would turn the connection between a mathematical statement and its proof into a purely external matter (Shanker 1990: 221ff). This strongly suggests that the first incompleteness theorem should be formulated in syntactical fashion, without reference or commitment to truth, as stating that every formal system $S$, if consistent and when elementary number theory is taken as its domain, contains a formula A expressing a proposition $A$ of elementary number theory such that neither A nor its negation ¬A, expressing $\neg A$, is provable in $S$.

So, to begin with: May we or may we not claim that Gödel's undecidable formula is true *simpliciter*, or true *tout court*, as distinct from true beyond recognition, either effectively or in principle, by us, humans? When giving an informal sketch of the main idea of the proof in the first section of his 1931 paper, Gödel says that if the proposition $[R(q);q]$ were provable, it would also be correct [*richtig*] and that, in that case, $\overline{\text{Bew}}\ [R(q);q]$ would hold [*würde gelten*], "which contradicts the assumption" (Gödel op. cit.: [175] 149).[4] If, on the other hand, the negation of

[$R(q);q$] were provable, Bew [$R(q);q$] would hold. Then both [$R(q);q$] and its negation would be provable, "which again is impossible." He then concludes this introductory section by saying that "[f]rom the remark that [$R(q);q$] says about itself that it is not provable, it follows at once that [$R(q);q$] is [correct] [*richtig ist*], for [$R(q);q$] *is* indeed unprovable (being undecidable). Thus the proposition that is undecidable *in the system PM* still was decided by metamathematical considerations" (Gödel op. cit.: [176] 151).

Gödel does not use the German *wahr* in this instance. Jean van Heijenoort resorts to the English equivalent of that German word in his translation (which I have departed from here on purpose), and Kleene, in his presentation, also claims that the formula A is "unprovable, hence *true* [emphasis mine]" (Kleene 1986: 128).[5] Is it sensible, then, to claim that truth forces its way into the Gödelian picture—which after all is entirely in terms of a proposition being *correct* and in terms of a claim to the effect that $x$ is not a provable formula (or, better, in terms of a claim to the effect that a natural number $q$ belongs to a class $K$ of natural numbers, with $K$ defined in terms of non-provability) *holding—only* because of van Heijenoort's translation, and that Kleene's presentation is, likewise, flawed, or at least anomalous in this respect?[6]

It is not, even though there is *no* notion of "correctness" or of "holding" to be *mis*translated in THEOREM VI itself:

> For every ω-consistent recursive class $k$ of FORMULAS there are recursive CLASS SIGNS $r$ such that neither $v$ Gen $r$ nor Neg ($v$ Gen $r$) belongs to Flg($k$) (where $v$ is the FREE VARIABLE of $r$).

Although Gödel's formulations do not involve a direct or explicit claim to the effect that the undecidable formula is true, or true without any further proviso or qualification, but only a claim to the effect that, for all $x$, $x$ is not the Gödel number of a proof of it, it can hardly be maintained that the formula which is undecidable modulo the consistency and ω-consistency of $P$, and which states that it is neither provable nor refutable in the system, may *not* be a truth bearer (and thus, may *not* be true *simpliciter* and, a fortiori true *and* undecidable).

As far as the informal presentation is concerned, the undecidable formula *truly* says of itself that it is not provable for it *is* indeed not provable. Is the situation any different when, instead of referring to the undecidable formula by means of its metamathematical description [$R(q);q$], we refer to it by means of its Gödel number once we have determined the number $q$, i.e., by the expression "17 Gen $r$" ("$x$ Gen $y$" denoting the 15th number theoretic function proven to be (primitive) recursive)? Undeniably, Gödel concludes his proof of the first incompleteness theorem by saying that "17 Gen $r$ is therefore undecidable on the basis of $k$, which proves THEOREM VI" (Gödel op. cit.: [189] 177) and *not* by saying that "17 Gen $r$ is therefore true and undecidable on the basis of $k$, which proves THEOREM VI." The undecidable formula nevertheless *truly* claims that 17 Gen $r$ is not $k$-PROVABLE and that Neg (17 Gen $r$) is, likewise, not $k$-PROVABLE.

Of course, Gödel remarks that "the purpose of carrying out the […] proof with full precision […] is, among other things, to replace the [assumption that every

provable formula is *true* [[my emphasis]] in the interpretation considered] by a purely formal and much weaker one" (Gödel op. cit. : [176] 151). Kleene might be right to explain that we assumed that only true formulas are provable in *S* to the extent that this assures us that "the formulas have clear meanings" (Kleene loc. cit.), but this is so provided that only true formulas are provable *no matter how one refers to them*. The important point here is not about meaning, or limpidity of whatever must be grasped. What counts is that it must not matter in this respect that 17 Gen *r* is a sentential formula whose undecidability has to be stated as being about this particular SENTENTIAL FORMULA. Kleene is certainly right to remind us at this point that the informal concept of truth was not "commonly accepted as a definite mathematical notion, especially for systems like [*Principia Mathematica*] or Zermelo-Fraenkel set theory" (Kleene loc. cit.). It nevertheless remains that whether we are dealing with a metamathematical description of the undecidable proposition or with the undecidable proposition itself (see Gödel op. cit.: [175n13] 149n13 for the distinction), we are indeed in a situation where contradicting (simple) consistency would yield the *falsity* of [$R(q)$; $q$] (and, likewise, the falsity of 17 Gen *r*), and where contradicting ω-consistency would similarly yield the *falsity* of its negation (and, likewise, the falsity of the negation of 17 Gen *r*), as well as the falsity of an assertion to the effect that the negation of these formulas is, respectively, provable and *k*-PROVABLE.

The assertion of its own unprovability and irrefutability qualifies A for the status of truth bearer and the price one must pay for the jettisoning of consistency and ω-consistency is indeed, by parity, that of its falsity. Bivalence, here, is not the issue. The issue is whether the purely formal and much weaker assumption that has replaced the informal and stronger assumption that every provable formula is true has disposed of our problem, or dissolved it into thin air, and my point here is that it has not.[7]

## 3. The Logical Constants

Another set of remarks that turn out to be relevant for the forthcoming argument concerns Gödel's conception that "intuitionistic logic, as far as the calculus of propositions and of quantification is concerned, turns out to be rather a renaming and reinterpretation than a radical change in classical logic" (Gödel [1941] 1995: [3] 190). Gödel defended this view after the publication of the undecidability results, first in a paper given at Karl Menger's colloquium in Vienna in 1932 (Gödel [1933] 1986), then in a lecture delivered at Yale in April 1941, from which the last quote is taken. He then discussed the view with respect to the issue of the use of abstract intuitionistic proofs in the explanation of the intuitionistic logical constants in the *Dialectica* paper from 1958 (Gödel [1958] 1990; see also Gödel [1972a] 1990). What is of interest to us here is that Gödel, building on results by Glivenko, showed that the classical propositional calculus is a subsystem of the intuitionistic

propositional calculus and that every valid classical formula also holds in Heyting's propositional calculus provided that we translate the classical "notions" or "terms" (Gödel's words), or "operators" (Kleene's word), i.e., the classical constants, into the intuitionistic ones (Glivenko 1929; Gödel [1933] 1986).

In particular, Gödel defends the somewhat surprising claim that the law of excluded middle is intuitionistically acceptable. In classical propositional logic, $\sim$ being the classical negation sign, the formula $p \vee \sim p$ is a tautology. According to Gödel, although intuitionists reject this law for *their* notion of disjunction, one may nevertheless define *another* notion of disjunction in terms of the other primitive logical constants of *their* calculus so that, $\neg$ being the intuitionistic negation sign, $p \vee \neg p$ is *also* a tautology. It is sufficient, Gödel claims, to define, quite trivially, p $\vee$ q as $\neg(\neg p \bullet \neg q)$; p $\vee$ $\neg$p may then be translated, just as trivially, into $\neg(\neg p \bullet \neg\neg p)$ so that the law of excluded middle turns out to be a special case of the law of contradiction, which is, of course, intuitionistically valid (Gödel [1941] 1995: [2] 190).[8]

This noticeably overlooks the fact that intuitionists will want to reject classical deduction rules such as double negation elimination and classical tautologies such as excluded middle precisely because they contend that such rules and tautologies would allow us to draw *illegitimate* inferences (from judgments to judgments) on the basis of relations of *alleged* logical consequence holding between antecedent propositions and consequent propositions. What they will object to, while still holding on to the law of contradiction, is the very idea that either $p$ or its negation is true, or holds, whether or not we could either be able to decide the truth-value of $p$ or that of its negation. What is at stake here is the contention that the truth of $p$, or that of its negation, is independent and indeed cut off from all links with our ability to decide the matter one way or the other. Moreover, intuitionists require that the assertion of a negated proposition be justified by a *reductio ad absurdum* of the supposition that we could obtain a proof of the proposition, or of the supposition that the means of obtaining such a proof are, at least in principle, at our disposal. This strongly suggests that neither disjunction nor negation may be translated in the way suggested.

In yet other words, the following translation manual:

CLASSICAL LOGICAL CONSTANTS

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| $\sim p$ | $p \rightarrow q$ | $p \vee q$ | $p\&q$ |

INTUITIONISTIC LOGICAL CONSTANTS

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| $\neg p$ | $\neg(p \bullet \neg q)$ | $\neg(\neg p \bullet \neg q)$ | $p \bullet q$ |

overlooks the fact that, although no special symbol for intuitionistic disjunction is thereby introduced, the claim to the effect that $p \vee \neg p$ is valid is now tantamount to the claim that we have a constructive proof of $p \vee \sim p$, or at least the means of obtaining one. But the claim that we have a constructive proof of $p \vee \sim p$ is

acceptable only provided that we have either a proof of $p$ or a proof of $\sim p$ and this is definitely *not* what we have with classical disjunction. It seems, as a matter of fact, that what we have now with the translation manual is rather a claim to the effect that, independently of our knowledge, and perhaps indeed unbeknownst to us, either there is a proof of $p$ or there is a proof of its negation. Or perhaps what we have is a claim to the effect that either we have a proof of $p$ or we have a proof that we shall never have a proof of its negation, i.e., a proof of the double negation of $p$. But, clearly, the second term of this disjunction must be rejected by the intuitionist for a proof of the double negation of $p$ *doesn't* amount, intuitionistically, to a proof of $p$ (see Endnote 8).

In other words, we are surreptitiously appealing to an objective realm of proofs which is disconnected and indeed "cut off from all links with the reflecting subject," to borrow Bernays' apt phrase in his description of platonism in the philosophy of mathematics (Bernays [1935] 1983: 259; see also Bernays op. cit.: 267). In doing so, we are appealing to a notion of proof that is obviously *not* faithful to the constructive standpoint.[9] If this is the case, the Glivenko-Gödel translation manual leaves both the classical logician and the intuitionistic logician unsatisfied, precisely because, most vividly in the case of excluded middle, both will judge that the meaning imposed upon disjunction and negation by way of the translation manual is arbitrary.

Although it may seem at first blush that the validity of excluded middle does not depend upon any peculiarity in the interpretation of disjunction that intuitionists would object to on the ground that the translation manual would propose a definition of disjunction "in terms of *their* [emphasis mine] other primitive logical symbols" (Gödel [1941] 1995: [1–2] 190), i.e., in terms of intuitionistic conjunction and negation, we are indeed in the situation that Dummett describes thus:

> The failure of the law of excluded middle is often explained by the different meaning of intuitionistic disjunction: a proof of $A \vee B$ is a proof either of $A$ or of $B$, and hence a claim to have proved $A \vee \neg A$ amounts to a claim either to have proved $A$ or to have proved $\neg A$. Such an explanation of the matter is correct as far as it goes, but it will naturally leave a platonist with the feeling that the meaning imposed upon $\vee$ is arbitrary: on any view on which either $A$ or $\neg A$ must be true, irrespective of whether we can prove it, to repudiate that sense of $\vee$ in which we can assert $A \vee \neg A$ a priori is to deny ourselves the means of expressing what we are able to apprehend.
>
> Dummett (1977: 18)

There is in particular, at this fundamental level of primitive logical laws, something that seems to undermine the translation claim inherited from Glivenko and that Dummett does not underline in his objection, namely a disagreement about the logical form that a proper *reductio* should take and which directly concerns negation. It must be remarked, and indeed stressed in this instance, that a classical *reductio* is *not* equivalent to an intuitionistic one and that each yields a particular form of negation, so that $\sim$ may not, after all, be translated into $\neg$. The intuitionist's rationale for the rejection of *both* $p \vee \sim p$ and $p \vee \neg p$ *indiscriminately* under the proposed translation scheme is that neither formulation of excluded middle amounts

to the claim that we have either obtained a proof of $p$ or a proof of its negation, or that we are in a position to obtain either. So the distinction between the allegedly objectionable $p \vee {\sim}p$ and the quite acceptable $p \vee \neg p$ modulo the Glivenko translation turns out to be invidious: It points quite unfairly to a difference that, as a matter of fact, does not exist at all.

Let me now take stock of what has been established in Sects. 2 and 3. The first point is that A is a truth bearer; the second is that the validity of a logical law should not depend on any peculiarity in the assignment of a meaning to the main constant occurring in a logically valid statement or formula that prevails as a law. The first point matters because it licenses us to give a semantic formulation of Gödel's first incompleteness theorem. The second matters because, as Gödel insists, the purpose of carrying out of the proof of THEOREM VI with full precision is to replace the assumption that every provable formula is true in the interpretation by a *weaker* one. There is no doubt that the precision is indeed provided by the proof. The point here is that there should not be any arbitrariness in our determining whether or not the notion of truth involved in the claim that A truly asserts its own unprovability and irrefutability is itself constrained by provability or, on the contrary, unconstrained so that the truth of that claim could after all remain beyond recognizability by us, humans. As the following section should make clear, the notion of truth involved in the recognition of the truth of the proposition that asserts its own unprovability with no faulty circularity is not arbitrarily construed. If it were, we would have begged the question. With these points in mind, let me now turn to the semantic formulation.

## 4. Gödel's Theorem (I): A Semantic Formulation

The main outline of the semantic formulation of Gödel's proof may be given in the following way.[10] The proof exhibits an elementary formula which is finitary in Hilbert's sense and proven to be both unprovable and irrefutable in $P$, as the result of the following steps:

1. A formula A is constructed by diagonalization, which asserts its own unprovability in $P$.

2a. The consistency of $P$ being taken for granted, it is proven that A is unprovable in $P$.

2b. Since A asserts its own unprovability in $P$, A is true.

3. The ω-consistency of $P$ being taken for granted, it is proven that A is irrefutable in P.

4. A is proven to be undecidable in $P$.

5. A is true and undecidable in $P$.

Our warrant for (5) is that Gödel's formula A is both recognized to be true (at step (2b)) and proven to be undecidable (at step (4)) since it is both proven to be unprovable (at step (2a)) and proven to be irrefutable (at step (3)). Our question from Sect. 1 is whether truth may transcend its recognizability by us, humans. We have already remarked how large that question is; not just large but truly inordinate. A sense of balance or proportion might perhaps be restored if we answer two questions that are related to it according to the description of the doctrine Dummett has dubbed "realism."

The first question is whether our understanding of the meaning of A amounts to a knowledge of its truth conditions.[11] In particular, are we able to manifest, make plain, or display that we do possess that knowledge? We must answer this question in the positive since the truth of A is *recognized* as obtaining at step (2b).[12] We do have a means at our disposal to find out that the truth conditions of A are satisfied and are able to make that knowledge manifest by proving the unprovability of A under the assumption that $P$ is consistent (step (2a)) and by concluding that A is true (step 2b)). Our answer is positive because, as Gödel points out (in relation to the Epimenides paradox):

> [we] can construct propositions which make statements about themselves, and, in fact, these are arithmetic propositions which involve only recursively defined functions, and therefore are undoubtedly meaningful statements. It is even possible, for any metamathematical property $f$ which can be expressed in the system, to construct a proposition which says of itself that it has this property.
>
> Gödel [1934] (1986: [21] 362–363)

It is therefore possible, for any predicate $F$ of the language $\mathrm{L}_P$ of $P$ expressing in $P$ a given metamathematical property, to construct by diagonalization a formula A of $\mathrm{L}_P$ which asserts of itself that it possesses that property. If we note the Gödel number of that formula with the symbol "$\langle A \rangle$," then, for every predicate $F$ of $\mathrm{L}_P$, there exists a formula such that $\mathrm{A} \Leftrightarrow F(\langle A \rangle)$.

Let us choose as a metamathematical property the property of non-provability in $P$, expressed in $P$ by the predicate "non-$\mathrm{Pr}_P$." We may then construct a formula A which asserts its own unprovability in $P$, such that $\mathrm{A} \Leftrightarrow \text{non-}\mathrm{Pr}_P(\langle A \rangle)$.

Once that first step is accomplished, we may proceed to step (2a) and distinguish the following substeps leading to (2b).

If A were provable in $P$, then:

2a1. $\mathrm{Pr}_P(\langle A \rangle)$ would be true in $P$ and, therefore, provable in $P$, and
2a2. non-$\mathrm{Pr}_P(\langle A \rangle)$ would be provable in $P$, since A and non-$\mathrm{Pr}_P(\langle A \rangle)$ are logically equivalent.
2a3. $P$ would therefore be inconsistent.
2a4. Under the assumption that $P$ is consistent, A is therefore unprovable in $P$.

Gödel notes in this respect that:

> Contrary to appearances, such a proposition [which says of itself that it is not provable] involves no faulty circularity, for initially it [only] asserts that a certain well-defined formula (namely the $q$th formula in the lexicographic order by a certain substitution) is unprovable. Only subsequently (and so to speak by chance [*gewissermaßen zufällig*]) does it turn out that this formula is precisely the one by which the proposition itself was expressed.
>
> Gödel [1931] (1986: [176n15] 151n15)

We may then directly proceed to step (2b): Since A $\Leftrightarrow$ non-$\mathrm{Pr}_P(\langle A \rangle)$, A is true. It is thus clear that the question whether or not we know the truth conditions of A may—and indeed must—be answered in the positive by the time we reach (2b), for we can make it perfectly plain, by proceeding from step (2a1) to step (2b) that we know indeed that these truth conditions are satisfied. So it is possible for us, humans, to manifest (to borrow once again from Dummett's terminology) our knowledge of the truth conditions of a formula proven to be unprovable in a formal system obtained from Whitehead and Russell's *Principia*, without the ramification of the types, that takes the natural numbers as the lowest type and incorporates the usual Peano axioms.

As far as truth conditionality is concerned, we are not in a situation where it would be appropriate to eschew the twin notions of truth and truth conditions altogether. Since arithmetic propositions that involve only recursively defined functions are "undoubtedly meaningful statements" (Gödel [1934] 1986: [21] 362), there is no reason to jettison the truth conditionality principle, as applied to A. The meaning of A is indeed constituted by its truth conditions and our knowledge of that meaning amounts to a knowledge of these conditions. There is indeed something in virtue of which A is true, i.e., something in virtue of which its truth conditions are fulfilled, namely the proof that proceeds from (1) to (2b).

## 5. Gödel's Theorem (II): Two gaps

The second question related to the description of the doctrine Dummett has dubbed "realism" is whether we are in a case where truth transcends, one way or another, recognition by us, humans: either recognition in principle (in which case *theoretical or ideal recognizability* is at stake), or effective (in which case *actual or feasible recognition* is at stake). This, of course, takes us back to the vexed question we started with in Sect. 1, but we are now in a somewhat more comfortable position for we may after all get a purchase on that controversy. Gödel's proof is unambiguous in this respect: The elementary formula proven to be undecidable in $P$ given the consistency and ω-consistency of $P$ is true in a sense that may not offend either a

constructivist or Dummett's antirealist. As noted at the beginning of Sect. 2.1, Gödel thought it worthwhile to remark that his proof was unobjectionable from the intuitionistic standpoint. In particular, the proof does not allow one to conclude that the truth conditions of A transcend its justification conditions; it shows something quite different, namely that:

> […] our capacities for justification go beyond what is strictly speaking provable in a formal system: there exists, for each sufficiently rich formal system [i.e. such that the property "provable in the system" is expressible in the system], undecidable elementary statements that we nevertheless have cogent reasons to hold as true.
>
> Dubucs (1992: 57)[13]

In other words, Gödel's first incompleteness theorem does not show, either directly or indirectly—i.e., either with or without the help of a Princess Margaret Premise—that the extension of the predicate "true" is larger than the extension of the predicates "recognizable (in principle) as true" and "(effectively) recognized as true." The first gap would indeed be unacceptable from a constructivist or antirealist standpoint, and the second from a strict finitist one. What the proof of the theorem shows is that the extensions of the last two predicates are larger than the extension of the predicate "provable in *P*," which is quite another matter. What we have been able to acknowledge so far is that the truth conditions of A are transcendent with respect to its provability in *P*, but this offers nothing in the way of an admission of some "absolute" notion of recognition-transcendent truth, of some supreme notion, as it were, of *truth beyond all possible justification*, or even of the possibility thereof.

Two conclusions may thus be drawn.[14] The first is that there is no elementary formula whose truth could be undetectable in a formal system if we assume that system to be consistent (which we do). The most we are allowed to say is that there are elementary formulas whose truth remains algorithmically undetectable given the consistency of the system, and that amounts to a quite different claim. In the case in point, no algorithmic procedure may help us to conclude that A is true, but its truth is nevertheless acknowledgeable by us by means of a *reductio ad absurdum* of the supposition that it is provable in *P*, given that *P* is consistent. Unless we decree that the unavailability of an algorithmic procedure for deciding the truth-value of a formula is a criterion for the undetectability of its truth, there are no undetectable truths, or truths beyond all possible recognition in a formal system if that system is consistent. So the question is: Should we order the decree? Step (2b), from the previous section, strongly suggests that we may not even argue for this position (let alone decree that we may benefit from such a criterion). We must on the contrary distinguish the case of algorithmic undecidability from that of undetectability of truth-value *simpliciter* (or, better, in the case at hand, of undetectability of truth-value by any *non*-algorithmic means) when discussing the vexed question we started with.

The second conclusion to be drawn is thus that we must take into account a finer-grained distinction than the one we have been pondering over so far, i.e., one that contrasts:

A. The gap between what is true in the standard model for arithmetic and what is recognizable as true on the basis of cogent reasons

with

B. The gap between what is recognizable as true on the basis of cogent reasons and what is algorithmically recognizable as true in the standard model for arithmetic.

The first gap is filled by the proof of the unprovability of A and, a fortiori, by the (complete) proof of its undecidability. The second may *not* be filled, just because of the very same proof.

The question now is: What does this tell us about the vexed question we started with?

## 6. The Missing Princess Margaret Premise: Benacerraf's Assessment

When discussing the case of Gödel's incompleteness results, and of the first result in particular, Benacerraf drops more than a gentle hint at what the Princess Margaret Premise is, and indeed should be, with respect to the anti-mechanistic conclusion that we are not machines, especially when that conclusion is based on the "libertarian arguments to the effect that our abilities transcend those of any machine, outstrip in truth-power any formal system, i.e. do not constitute or cannot be adequately represented as an i.e. set of sentences etc." (Benacerraf 1996b: 42–43). The proper PMP one would have to add to the formal result in order to get the desired conclusion is that:

> There is something human mathematicians can do that no machine can do — for any (theorem proving) machine, find its Gödel number and, given its Gödel number, prove its Gödel sentence (something it manifestly cannot do).
>
> Benacerraf op. cit.: 31

Although the finer-grained distinction does allow us to conclude that there is something human mathematicians can do that no machine can do, it does *not* thereby support the much stronger view that *we are not machines*. For the sake of simplicity, let us call the claim that we must take into account the distinction between (A) and (B), along with the remarks about gap-filling (and the impossibility of gap-filling) offered at the very end of Sect. 5, the "two gaps thesis." We are *not* in a situation where we could avail ourselves to a specific instance of an argument of the form:

$$(\mathrm{C}) \quad (\textit{Metamathematical result}, \textit{PMP}) \vdash \textit{Conclusion}$$

with:

$$(\mathrm{C}^*) \quad (\textsc{Theorem} \text{ VI}, \text{Two gaps thesis}) \vdash \textit{We are not machines}$$

in the role of the desired specific instance.

If we were arguing in this way, we would indeed be deriving an unwarranted philosophical conclusion. Benacerraf knows this, of course, and remarks in this respect that:

> The [first incompleteness] theorem heralds what its name suggests: an incompleteness in formalized arithmetic. However hard we may squeeze, we can not extract from it the thought that, although we once believed we had a concept of arithmetic truth, now that we see that the sentences true in arithmetic cannot be exactly those corralled by any plausible formal "proof" procedure (if bivalence is also to be preserved), we must concede that we did not. The incompleteness that Gödel demonstrated, the incompleteness of the First Incompleteness Theorem, was shown to exist in the calculus, not in our conception. Not that one couldn't be lurking in our conception as well — of course one could — but that just hasn't been shown; […] to go that extra mile requires a Princess Margaret Premise and a separate argument to support it.
>
> Benacerraf op. cit.: 42

The two gaps thesis supports a much weaker view. It is therefore crucial, in order to identify that view correctly, to avoid either crediting the human mind with cognitive capacities it does not have, or denying that a Turing machine associated with the relevant formal system of arithmetic may perform tasks that it is, after all, clearly able to perform. In particular, it might be objected that although we are able to provide a justification for A, given that the formula correctly expresses its own unprovability, the availability of such a justification through a non-mechanizable step does not thereby establish that we are not a Turing machine, but only that we are not a Turing machine *associated with* P. After all, since $P$ is taken to be consistent and proves $Cons(P) \rightarrow \mathrm{A}$, the system $P^* = P \cup \{Cons(P)\}$ also proves A. A machine that would enumerate the arithmetical theorems of $P^*$ would indeed be able to generate A. We would not, therefore, be in a case where some arithmetical truth has been omitted and where the human mind would thereby be cognitively "superior" to the machine associated with $P^*$.

The anti-mechanist or libertarian might object at this point that she is not arguing that there exists an elementary formula of arithmetic whose truth a human mind is able to justify but that no consistent Turing machine will ever engender (or that such a machine will necessarily omit). She might wish to claim that her argument establishes the falsity of the general claim that Turing machines may, qua consistent, enumerate all the arithmetical truths for which (non-mechanical) human minds are able to find, albeit non-mechanically, a justification. She would then be proposing, in order to stand her ground, a case-by-case refutation of *each particular instance* of the general mechanistic claim. This would imply that the mechanism

involved in the libertarian claim to be rejected and the libertarian argument to be refuted amounts to the rather weak proposal that the mind is a machine, *but only given that it is established that, for each particular machine we may care to consider when assessing the mind-machine metaphysical identity thesis, the mind isn't* that *particular machine*. Some ω-inconsistencies would indeed be involved in such a conception of the non-mechanistic mind.[15]

Notice that Benacerraf, in Benacerraf (1967), seems to argue in favor of such a thesis: He infers from Gödel's incompleteness results that we are not Turing machines, *or at least that, if we were, we wouldn't be able to determine our own instruction tables*. This, of course, is a crucial proviso for it seems indeed to be a way of saying that the mind is a machine, albeit with the rather damaging caveat that it may be established, for each particular machine, that the mind is not *that* machine, just because the mind cannot acknowledge what its own Turing instructions are and is therefore clearly deficient in terms of self- or introspective knowledge.

We are left, then, with a position that draws as a conclusion from Gödel's first incompleteness result some indulgent form of libertarianism based on a weak notion of mechanism; a notion so weak indeed that it cannot do justice to the idea at the heart of libertarianism that our cognitive abilities transcend those of any Turing machine and, in particular, that such machines are unable to enumerate all the formulas whose truth a human mind or agent can acknowledge or recognize insofar as the mind or agent targets *the standard model for arithmetic*.

There must be another way to do justice to Benacerraf's important remark that Gödel's incompleteness should not be located in our *conception* of arithmetical truth and should safely remain where it belongs, i.e., in the *calculus* (or in the family of *calculi*) we have managed to devise. What is the view, then, which the two gaps thesis may support, so that one is neither "brandishing" the metamathematical result as the authority for some purely philosophical yet unwarranted conclusion (Benacerraf 1996b: 43), nor defending one that involves ω-inconsistency?

## 7. In Defence of the Proper Princess Margaret Premise

It looks like I have driven myself into a tight corner. I have claimed in Sect. 6 that there is something we can do that Turing machines cannot do, i.e., fill the first gap, and that this distinction supports a philosophical view which is both weaker than the strictly anti-mechanistic or libertarian view, and distinct from the one Benacerraf defends—or at least seems to defend—in Benacerraf 1967. Whatever its details might turn out to be, the philosophical conclusion we may draw must be true to Benacerraf's brief that the incompleteness is in the calculus and not in our conception of arithmetical truth. If the PMP is indeed one, the philosophical conclusion it yields must be consistent with the view that our conception of arithmetical

truth, or of what makes a statement of arithmetic true when it is true, should be the very notion we had before Gödel proved his incompleteness result, so that the "unformalized practice of mathematics" (Benacerraf 1996b:12) must escape unscathed from the Gödelian result. The point I wish to make with respect to the two gaps thesis being the proper PMP we need is that, contrary to what Benacerraf contends, something may *not* remain unscathed, namely the "implications […] regarding the very nature of […] ourselves as [the] practitioners [of mathematics]" (Benacerraf op. cit.: 14). Although we must reject the claim that our former conception of arithmetical truth was defective, despite the undeniable fact, established by Gödel's result, that "the sentences true in arithmetic cannot be exactly those corralled by any formal 'proof' procedure (if bivalence is also to be preserved)" (Benacerraf op. cit.: 42), we still must conclude that there is *a human cognitive capacity that transcends those of any Turing machine.* It is the conception of "ourselves," or of our minds, or of the scope and nature of our own cognitive capacities that must be amended so that the incompleteness is, strictly speaking, a property of the mathematical formalism, and nothing more.

Benacerraf claims that the argument that purports to show that we are

> subject to the same limitations that have been proved to hold of the formal languages and systems that we study in metamathematics (or, in other cases, free from them) […], if it is to be at all probative […], must include a convincing demonstration of the relevant isomorphism (or lack thereof) between our own powers and the relevant features of the systems.
>
> Benacerraf op. cit.: 12

The lack of isomorphism is manifested by our capacity to fill the gap between what is true in the standard model for arithmetic and what is recognizably true on the basis of cogent reasons, and this is precisely what allows us to conclude that we are not machines, albeit neither in the strong metaphysical sense that we are not strictly speaking identical to any Turing machine, nor in the weaker sense that we are not any particular machine we may care to consider, given our ignorance, in each and every particular putative case, of our own instruction tables. The crucial point here is that Gödel's incompleteness proof, by showing that the first gap is filled, *thereby shows that the second may not be filled.* So our PMP relies crucially on the claim that we have the cognitive capacity to target *the* standard model for arithmetic, for it is this very capacity which is responsible for our acknowledgement of the truth of A, i.e., the capacity that so troubles Shanker (see, supra, Sect. 2).

To go back to the first Benacerrafian methodological point that Morton and Stich remind us of (see Morton and Stich 1996: 5) and that I have mentioned at the very beginning of this paper in the first paragraph of Sect. 1, it would be strange indeed, perhaps even mystifying, if nothing philosophical could be "proven" or argued in this regard, so that nothing more than the price of the philosophical point about ourselves, or our minds, or our cognitive capacity to target the standard model of arithmetic, would be known. After all, if we have shown the price of what must be proven, we know the price, and if we know the price, the only thing that could prevent us from coming by the conclusion is some unfortunate lack of funds. But how could that be? If we have determined which argument would yield a

philosophical view when appended to an established metamathematical result, then we do, *de facto*, have that argument in favor of the purely philosophical conclusion, although not necessarily an independent one for each of its premises.[16]

Benacerraf remarks, of course, that the metamathematical result *alone* yields a *weaker* purely philosophical conclusion (Benacerraf op. cit: 43), so that if (C), then:

$$(\mathrm{C}') \quad \textit{Metamathematical result} \vdash (\textit{PMP} \rightarrow \textit{Conclusion})$$

in which case

$$(\mathrm{C}'^{*}) \ (\textsc{Theorem VI}) \vdash (\text{Two gaps thesis} \rightarrow \textit{We are not machines})$$

would be the specific desired libertarian instance.

I have argued here, though, in favor of a different and somewhat weaker claim. We do need, of course, independent grounds for that claim, grounds which would give it a bona fide proper content: We do need to say more on the nature of these "cogent" reasons. I will not be arguing for such independent grounds here. What may be stressed, though, is that since A is a genuine truth bearer and the notion of truth involved in the claim that A truly asserts its own unprovability is neither guilty of faulty circularity nor arbitrarily construed, Gödel's proof, together with the two gaps thesis, yields a claim to the effect that we are free of one limitation that has been proved to hold for *P*, where *P* is the system obtained from *Principia Mathematica*, without the ramification of the types, taking the natural numbers as the lowest type and adding their usual Peano axioms.

We cannot guarantee any public display of our non-mechanical knowledge that the truth conditions of A obtain, given that *P* is consistent, over and above what the argument sketched in Sect. 4 imparts with steps (1)–(5) (and steps (2a1)–(2a4)). This, obviously, is a defect for which the semantic formulation argued for in Sect. 2 is fully responsible. Shanker, or anyone convinced by his skepticism, may rightly object that since the reference to the standard model may not be eliminated from our argument, and since no recursively axiomatizable class of formulas allows us to give a proper definition of such a standard model, the argument is wanting unless it may be shown, once again on independent grounds, that our grasp of the standard model need *not* rely on the existence of a language fit to describe it, say second order Peano arithmetic, or some other.[17]

What I would like to stress here in way of a conclusion is that, perhaps more than truth proper, the crucial philosophical dimension of the semantic reading of Gödel's first incompleteness theorem is bivalence, as the remarks of Sect. 3 about the meaning of the logical constants in relation to the validity of logical laws, and the quote, in this section, to the effect that arithmetic truth does not quite match arithmetic provability if bivalence is to be secured (Benacerraf 1996b: 42) clearly indicate. I shall now turn to this question.

## 8. Concluding Remarks About the Larger Picture

I promised in Sect. 1 to explain why it is important not to loose sight of the scope of the general question we started with in spite of the fact that it might appear too general, unspecific and, for this very reason, nothing less than ill-defined. There is no doubt that the question whether truth may go beyond recognizability in principle is a crucial philosophical question. The problem, rather, if the gist of the argument that unfolds from the beginning of Sect. 4 to the end of Sect. 7 stands up to critical examination, is whether we should care about that larger picture at all in the specific case of the Gödelian metamathematical result we have taken into consideration.

The large question has gained momentum and even currency in great part because of Dummett's persistent claim that it lies at the core of the realism versus antirealism debate, conceived as a debate whose genuine content is semantic in nature (hence the reference to those "Dummettian terms of transcendence, truth, recognizability and the in principle vs. effective distinction" in the opening paragraphs of Sect. 1). Part of the advantage of being at this center is that it discloses the non-metaphorical content of the time-honored metaphysical issue of the existence and, should existence be granted, of the independence of objects of some given kind, or sort, or class; typically, abstract objects such as numbers, but also, say, colors or values. Since semantics is a stake, the question of bivalence is crucial; not just of bivalence, but of semantic principles in general. More specifically, what is at stake is the question of their relation to logical laws. How is, say, bivalence related to excluded middle, or stability to double negation elimination? One idea is that the semantic principles justify the logical laws. So the question is: Should the semantic principles be accepted?

Benacerraf rejects the idea that since

> (bivalent) truth outruns formal provability, for any consistent formal system of Arithmetic, the concept (of bivalent truth) is inherently flawed and must be replaced with a concept of truth in arithmetic that is more closely tailored to our ability as provers. This, of course, then splinters into countless possibilities, depending on how bounded we believe our "proving" abilities to be.
>
> Benacerraf (1996b: 30)

The splintering might well be curtailed so that, after all, only two main possibilities remain, or at least two kinds thereof, depending on whether one rests content with in principle formal provability or wishes to insist that effective formal provability, guaranteed to be humanly feasible, must be secured in order for the principle of bivalence to be acceptable.

Given what has been said before about the implications of Gödel's result given some proper PMP, this conception of the issue at stake strongly suggests that we are facing a dilemma: Either we must conclude that we are somehow free of the shackles built in the mechanistic view because we are endowed with super-mechanical powers, or truth must be constrained by some epistemic notion such as provability (either in principle, which would satisfy Dummett's antirealist,

or effective, along strict finitist lines). In other words, it is as if we should choose between two conclusions to be derived, provided one may be derived at all with the aid of some bona fide PMP: either the adoption of a strong metaphysical libertarian thesis, or the replacement of bivalent truth conditions by conditions such that it must be guaranteed, by the very nature of the case, that we, humans, are able to recognize that they obtain when they do (or at least that we are able to put ourselves in the position to activate the appropriate recognitional capacities to that effect).

It is, I suspect, because of such a dilemma (either libertarianism or the death of bivalence) that Benacerraf takes Putnam to task for locating the incompleteness in our conception on the ground of an endorsement of a "heavily 'epistemic' notion of truth" (Benacerraf op. cit.: 44). Benacerraf's rationale for rejecting Putnam's negative conclusion regarding bivalence is that epistemic considerations, as applied to truth—and hence to the issue of the unrestricted appeal to the semantic principle—rather than offering reasons to be restrictive about what we are able to understand "point the opposite way." His justification for this position is that he

> [takes] "epistemic" considerations to include ones of meaning, but without accepting as axiomatic that these are ineluctably entwined with a semantics of "assertibility conditions," as opposed to a "truth-conditional" semantics, or even some other, as yet undreamt-of, kind.
>
> Benacerraf op. cit.: 44

It might seem odd, once epistemic considerations have been taken to include those of meaning, especially with respect to the issue of grasp or understanding, to claim that the latter are divorced from considerations pertaining to assertibility conditions. It is odd, or course, not because one would be an antirealist in Dummett's preferred sense, but because our cognitive limitations as "provers," i.e., those that play a key role in the PMP, come in a twosome, along with our *non*-mechanistic ability to fill the gap between what is true in the standard model and what is recognizable as true on the basis of cogent reasons (provided we are able to target the standard model and that the targeting need not rely on a language fit to give a bona fide description of it). These limitations come hand in hand with, as it were, a positive aptitude or capacity. The reason why we are not Turing machines after all is that being such a machine requires that there be a finite collection of instructions, each instruction calling for atomic operations to be performed under certain given conditions. The recognition of such a cognitive limitation or failure is therefore crucial, in Benacerraf's 1967 moderate anti-mechanistic conclusion. This is even more so in the argument I have sketched: There is a bound to our purely mechanical proving abilities, and a corresponding release, as it were, of our non-mechanistically characterizable bona fide recognitional capacities with respect to truth.

Putnam, quoted by Benacerraf, claims that since nothing epistemic may help us "explain the truth-value of […] undecidable statements, precisely because they *are* undecidable," i.e., no matter how hard we constrain the notion of arithmetic truth, we should refrain from attaching any "metaphysical weight to the principle of bivalence," a principle which would have us believe that these statements are either

true or false although "their truth-value cannot be decided on the basis of the axioms we presently accept" (Benacerraf 1996b: 44).

What I have been urging here, the acceptance of axioms notwithstanding, and the crucial matter of the targeting of the standard model pending, is to leave bivalence alone and to draw from a modest PMP an equally modest conclusion to the effect that there is a least one capacity that the anti-libertarian or hard-nosed computationalist is unable to account for.

## Notes

1. Benacerraf reports the following parable of the Cohens and Princess Margaret (Benacerraf 1996b: 9–10). The Cohens want the right spouse for their son. When they accept somewhat reluctantly the goy Princess Margaret as the right girl for Abie on the basis of the staggering advantages their future grandchildren would benefit from (being heirs to the throne of England, etc.), only *half* the shatchen's job is done. Abie still has to marry Princess Margaret. Without the Cohen's acceptance of the marriage broker's final offer, no result could be obtained, but more is needed nevertheless for the job to be rounded off. The Cohen's reluctant acceptance after the turning down of so many proposals is what Benacerraf mischievously dubs "the easy part"; what is needed now is the extra move which will bring the broker's efforts to its expected conclusion, i.e., a marriage settlement.
   In the same way, once you have obtained your metamathematical result and the first half of the job is done, you still need some extra premise—the so-called Princess Margaret Premise—along with an independent argument in its favor, to get the analogue of the settlement, i.e., the desired philosophical conclusion that you may not obtain directly from the metamathematical result.
2. An important part of what follows draws from Pataut (1998). That paper did not address Benacerraf's points about the PMP one must fall back on in order to draw philosophical conclusions from Gödel's first incompleteness theorem, although it did address the truth *vs.* recognition of truth issue. My purpose here is to address as explicitly as possible the points Benacerraf makes with respect to this opposition, especially the point about the demonstrated incompleteness being shown to exist *in the calculus* and not *in our conception* (see Benacerraf 1996b: Sect. 4.2, pp. 42–44). It is essential to avoid both the mathematical fallacy of deducing philosophical conclusions directly from formal results of a metamathematical kind, and the philosophical blunder of mistaking some irrelevant extra premise for the genuine Princess Margaret one.
3. Boolos has proposed a non-constructive proof in Boolos (1989). His proof, just like Gödel's, establishes the existence of an undecidable statement of arithmetic, but, unlike Gödel's, it does not provide an effective procedure for producing it. Let a correct algorithm $M$ be an algorithm which may not list a false statement of arithmetic. A truth omitted by $M$ is a true statement of arithmetic not listed by $M$. Boolos's proof establishes the existence of such a statement, but the statement is recognized as true classically and not constructively.

4. *Richtig* may also be rendered as "right," and *würde gelten* as "would be valid," or "would obtain."
5. A states that every natural number $x$ is not the Gödel number of a proof, in $S$, of a formula that turns out to be A.
6. Kleene notes that Gödel approved van Heijenoort's translation of his 1931 paper for the then forthcoming volume van Heijenoort was editing (van Heijenoort 1967), and that the translation was accommodated in many places to Gödel's own wishes (Kleene 1986: 141).
7. Gödel notes that "the true reason for the incompleteness inherent in all formal systems of mathematics is that the formation of ever higher types can be continued into the transfinite, while in any formal system at most denumerably many of them are available" (Gödel [1931] 1986: [191] 181, note 48a).
   Kleene's reading of this remark is that Gödel implicity defends the view that the adjunction of higher types "permits one to define the notion of truth for that system, then to show that all its provable sentences are true and hence to decide the sentence shown in THEOREM VI to be undecidable in the system" (Kleene 1986: 135). In that case, then, the sentence (or formula) whose undecidability has been decided thanks to the adjunction of higher types is indeed a bona fide true *and* undecidable sentence (or formula), as suggested toward the beginning of this section but, in that instance, without the recourse to types.
8. $\neg(\neg p \bullet \neg\neg p)$ is a case of the law of contradiction just in case double negation may be eliminated, something an intuitionist will *not* grant.
9. This, of course, must be judged against Gödel's complaint that, it is "doubtful whether the intuitionists have really remained faithful to their constructive standpoint in setting up their logic" and, worse, that "the notion of an intuitionistically correct proof or constructive proof lacks the desirable precision" to the point that "it furnishes itself a counter example against its own admissibility, insofar as it is doubtful whether a proof utilizing this notion of a constructive proof is constructive or not" (Gödel [1941] 1995: [3–4] 190).
10. It is a much too simple-minded formulation that does not do justice to essential features of Gödel's proof such as the Gödel numbering of the formal objects, the notion of numeralwise expressibility, and the constructively defined notion of the class of primitive recursive functions. Its one and only purpose is to focus on the relation between arithmetical truth and formal proof procedures in formalized arithmetic.
11. For the idea that A has a meaning, or is meaningful, see Sect. 2 and Kleene's remark in Kleene (1986: 128), already quoted in that section.
12. The claim should be qualified. At steps (2a)–(2b), we manifest our knowledge that *if* $P$ is consistent, then A is unprovable in $P$ and therefore true. There remains the further problem of knowing how we could know that $P$ is consistent and make that knowledge manifest. I have focused here on the consequent of the conditional, but it is of course established by Gödel's second incompleteness theorem (Theorem XI in Gödel [1931] 1986) that a proof of the

consistency of *P* may not be obtained in *P*. What we have here, therefore, strictly speaking, is only a *partial* manifestation of our knowledge of the truth conditions of A.

13. My translation from the French.
14. The claim should be qualified. The conclusions follow from the first part of Gödel's proof and are grounded on the steps of its semantic formulation up to (2b). I have not taken into consideration the proof of the unprovability of A (given the ω-consistency of *P*). Note that the very same conclusions would a fortiori be justified were the complete proof taken into consideration.
15. See Dubucs (1992: 75–76), to which I am very much indebted for these remarks.
16. See Benacerraf (1967, 1996b: 54, endnote 20) for a stern dismissal of arguments *without* PMPs, i.e., without "a significant injection of philosophical serum."
17. See Dummett ([1963] 1978) for a discussion of the manifestation or public warrant of our private but still commonly shared grasp of bona fide mathematical objects and formal proofs, in regard to Gödel's incompleteness result.

## References


Benacerraf, P. (1967). God, the Devil and Gödel. *The Monist*, 51 (January), 9–32.

Benacerraf, P. (1996b). What Mathematical Truth Could Not Be - I. In A. Morton and S. P. Stich (Eds.), *Benacerraf and his Critics* (pp. 9–59). Blackwell Publishers: Oxford and Cambridge, Mass.

Bernays, P. [1935] (1983). On platonism in mathematics. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics—Selected readings* (C. D. Parsons, Trans.), (2nd ed., pp. 258–271). Cambridge UP: Cambridge.

Boolos, G. (1989). A new proof of the Gödel incompleteness theorem. *Notices of the American Mathematical Society, 36*(4), 388–390.

Boolos, G. (1996). On the Proof of Frege's Theorem. In A. Morton and S. P. Stich (Eds.), *Benacerraf and his Critics* (pp. 143–159). Blackwell Publishers: Oxford and Cambridge, Mass.

Dubucs, J. (1992). Arguments gödéliens contre la psychologie computationnelle. *Travaux de logique n° 7 (juin 1992): Kurt Gödel, Actes du colloque, Neuchâtel, 13 et 14 juin 1991*, Denis Miéville (Éd.), 73–89, Centre de Recherches Sémiologiques, Université de Neuchâtel.

Dummett, M. A. E., Sir. [1963] (1978). The philosophical significance of Gödel's theorem. In *Truth and other enigmas* (1st ed., pp. 186–201). Duckworth: London.

Dummett, M. A. E., Sir. [with the assistance of Roberto Minio]. (1977). *Elements of intuitionism, Oxford Logic Guides n°2* (1st ed.). Clarendon Press: Oxford.

Glivenko, V. I. (Гливе́нко, Вале́рий Ива́нович) (1929). Sur quelques points de la logique de M. Brouwer. *Académie royale de Belgique, Bulletin de la classe des sciences, 15*(5), 183–188.

Gödel, K. [1931] (1986). Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I/On formally undecidable propositions of *Principia mathematica* and related systems I. In S. Feferman (Ed.-in-chief), *Collected Works: Vol. I Publications 1929–1936* (J. van Heijenoort, Trans.), pp. [173–198] 144–195. Oxford UP: Oxford.

Gödel, K. [1933] (1986). Zur intuitionistischen Arithmetik und Zahlentheorie/On intuitionistic arithmetic and number theory. In S. Feferman (Ed.-in-chief), *Collected Works*: *Vol. I Publications 1929–1936* (S. Bauer-Mengelberg & J. van Heijenoort, Trans.), pp. [34–38] 287–295. Oxford UP: Oxford.

Gödel, K. [1934] (1986). On undecidable propositions of formal mathematical systems [with the Postscriptum (3 June 1964)]. In S. Feferman (Ed.-in-chief), *Collected Works*: *Vol. I Publications 1929–1936*, pp. [1–27] 346–371. Oxford UP: Oxford.

Gödel, K. [1958] (1990). Über eine bisher noch nicht benütze Erweiterung des finiten Standpunktes/On a hitherto unutilized extension of the finitary standpoint. In S. Feferman (Ed.-in-chief), *Collected Works*: *Vol. II Publications 1938–1974* (S. Bauer-Mengelberg and J. van Heijenoort, Trans.), pp. [280–287] 241–251. Oxford UP: Oxford.

Gödel, K. [1941] (1995). In what sense is intuitionistic logic constructive? In S. Feferman (Ed.-in-chief), *Collected Works: Vol. III Unpublished essays and lectures*, pp. [1–30] 189–200. Oxford UP: Oxford.

Gödel, K. [1972a] (1990). On an extension of finitary mathematics which has not yet been used. In S. Feferman (Ed.-in-chief), *Collected Works*: *Vol. II Publications 1938–1974* (L. F. Boron, Trans., A. S. Troelstra, Revised), pp. 271–280. Oxford UP: Oxford.

Heijenoort, J. van (Ed.) (1967). *From Frege to Gödel: a source book in mathematical logic*, 1879–1931. Harvard UP: Cambridge, Mass.

Kleene, S. C. (1986). Introductory note to *1930b, 1931* and *1932b*. In K. Gödel *Collected Works*: *Vol. I Publications 1929–1936*, S. Feferman (Ed.-in-chief), pp. 126–141. Oxford UP: Oxford.

Morton, A. & Stich, S. P. (1996). Introduction. In A. Morton & S. P. Stich (Eds.), *Benacerraf and his Critics* (pp. 1–5). Blackwell Publishers. Oxford and Cambridge, Mass.

Pataut, F. (1998). Incompleteness, Constructivism and Truth. *Logic and Logical Philosophy, 6*, R. Lesko and H. Wansing (Eds.), 63-76.

Shanker, S. G. (1990). *Gödel's Theorem in focus*. Routledge: London.

# Part I
# Benacerraf's Dilemma

To say that mathematics is really logic in disguise merely pushes the problem off onto logic. If logic includes set theory, the problem is particularly difficult. I don't even know of an adequate answer to the question when limited to the propositional calculus and quantification theory. I suspect that the animal in question (the nature of mathematical truth) will turn out to be a many-headed monster; it will have to be slaughtered and appropriately butchered into pieces which are sufficiently manageable to lend themselves to fruitful dissection.

Benacerraf (1960: 255–256)

# Chapter 1
# McEvoy on Benacerraf's Problem and the Epistemic Role Puzzle

**Jody Azzouni**

## 1.1 Benacerraf's Problem

Benacerraf's problem is justly famous. It's had a major influence on the philosophy of mathematics right from its initial appearance,[1] an influence that continues up through the present moment. In its author's supernaturally elegant prose, it lays out a tension between the possibility of an epistemic access to abstracta and the apparent semantics (truth conditions) of mathematical statements about those entities. Given a causal construal of epistemic access, on the one hand, it seems that we can't have any epistemic access to the objects that our true mathematical statements must be about because those objects are causally inefficacious and causally insensitive; on the other hand, the mathematical truths in question are genuinely *about* those objects, and somehow we are adept at identifying some of the true mathematical statements and some of the false ones.

Benacerraf's problem long outlasted the faddish "causal theory of knowledge" that he originally couched it in terms of.

Field, among others, generalized Benacerraf's problem by writing:

> Benacerraf's challenge […] is to provide an account of the mechanisms that explains how our beliefs about these remote entities can so well reflect the facts about them. The idea is that if it appears in principle impossible to explain this, then that tends to undermine the belief in mathematical entities, despite whatever reason we might have for believing in them.
>
> Field (1989: 26)

J. Azzouni (✉)
Department of Philosophy, Tufts University, Medford, MA, USA
e-mail: jody.azzouni@tufts.edu



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_1

The challenge is now put in terms of mechanisms *of any sort*—they needn't be causal ones. What's important is that such mechanisms are required to explain—at least in principle—how our beliefs about these "remote" entities so well reflect the facts about them. So what's being indicated is the worry that because the entities in question are typically abstract (e.g., they're not in space and time, they're not causally efficacious or causally sensitive), no possible mechanism can—*even in principle*—foot the bill.

And indeed, one strand of the widespread philosophical response to Benacerraf's problem is precisely to supply an "in principle" mechanism for these remote entities by denying their purported *remoteness*. Maddy (Maddy 1990) suggests that at least some sets are *in* space and time, the set of the shoes that I'm wearing, for example, is claimed by her to be located where my shoes are at; she further suggests that I *see* this set each time I happen to gaze down at the shoes themselves. Field (similarly) suggests that points and spacetime regions aren't disagreeably remote because (after all) they're *in* spacetime; so too there's an *in principle* perceptual mechanism by which our beliefs about points and spacetime regions can so well reflect the facts about them.

Leaving aside the discussions of "in principle" mechanisms that (some) philosophers are so addicted to, Benacerraf's problem is recognizably a *modern* one, not visible to earlier philosophers concerned with mathematical entities. Plato, for example, can be seen as offering an *experiential*, if not perceptual, theory of the mechanisms involved. As disembodied souls we previously experienced the mathematical objects that some of us (in our current lives) nostalgically recollect. (*Those people* are called "mathematicians.") A later view, roughly attributable to Descartes, is that mathematical ideas are innately imprinted in our minds. The worry about such imprints being reliable remains—it's not solved by merely postulating that mathematical ideas are innate; it's solved (by Descartes) using the hypothesis that an "honest-broker" deity is the (ultimate) source of these ideas. It's that the metaphysics behind these earlier views are no longer respectable—so too, that certain epistemic accompaniments (e.g., "rational intuition") aren't respectable either—that's allowed Benacerraf's problem to emerge as a *problem* for contemporary philosophers of mathematics.

Meanwhile (elsewhere, in the sciences) a progressive understanding of the *actual* mechanisms of perception has emerged, especially in this and the last century. That is to say, the *actual mechanisms* behind our abilities to perceive those parts of the world that we can perceive are being uncovered. And this extends to our understanding of the instrumental mechanisms by which we've extended our sensory capacities; sophisticated instruments that gain us epistemic access to otherwise sensorily-remote parts of the universe (the very far away, the very small, for example) are themselves the subjects of intricate scientific studies.[2] Epistemic access (in empirical realms, anyway) is itself the subject of sophisticated science.

The remarkable fact about mathematics is that there is *nothing* corresponding to the scientific study of the *epistemic access* to its entities. There is, of course, serious cognitive-science studies of the (largely subpersonal) mental processes that occur when animals, children, and adults engage in mathematical ratiocination (e.g., when

they count some things that are near them)[3]; but in lieu of Cartesian deities, these studies of arithmetical competence hardly count as studies of *epistemic access*.

An important corollary of the emergence of studies of epistemic access is the systematic recognition of *epistemic artifacts*. Epistemic artifacts are the ways that our means of access to objects distort our impressions of the properties of those objects. In vision science, this takes the form of a systematic study of "optical illusions," in particular, the study of what it is about our visual capacities that allows such illusions to occur.[4] A corresponding study of our instrumental access to objects occurs in the sciences; we learn how the mechanisms by which an instrument allows us to learn about its target objects are limited in various ways, how they generate false impressions of the properties of the objects, and so on.[5] Understood broadly: these are studies of how we make *mistakes* about the objects we're epistemically accessing in particular ways; and it's our (scientific) *understanding* of the epistemic mechanisms we use to gain access to those objects that explains how those mistakes arise.

In the mathematical sciences, by contrast, mistakes are invariably "proof-theoretic ones." I understand "proof-theoretic" broadly: it can be a matter of our failure to execute a computation correctly, by actually writing down the wrong numerals,[6] for example; but it can also be a matter of conceptualizing a class of objects the wrong way. What it never involves, however, is that the mechanism of our epistemic access to the *abstracta* under study is misleading, that our means of epistemic access *to those objects* is itself creating problems. There is no study of the nature of such mechanisms; there is no science of such; there are no cases where we say, for example: *Rational intuition* often fails in such and such circumstances because…

Notice that these disanalogies with respect to epistemic access in these respective areas, mathematics and the various empirical realms, don't even remotely involve skeptical considerations—as philosophers generally construe those considerations, anyway. The point is an "internal" one about how the sciences, and the corresponding sciences of their various kinds of epistemic access to the world, have developed. I'll return to the purported relationship of these considerations about scientific studies of epistemic access to philosophical skepticism later in the paper.

## 1.2 The Epistemic Role Puzzle

In my *Metaphysical Myths, Mathematical Practice* (Azzouni 1994), I argued that the influence of Benacerraf's problem on the philosophy of mathematics literature was deeply *nefarious*: focus on *it* has obscured our view of the real philosophical issues and puzzles posed by standard mathematical practice. Attempting to establish this, however, involved emulating Benacerraf's achievement by posing a *different* philosophical puzzle, one that can be seen as derived (as it were) from Benacerraf's problem by *flipping* the angle of concern. Instead of worrying about how it is we're supposedly getting access to these "remote" objects, notice instead, as the

concluding paragraphs of the last section indicated, that mathematical objects play no epistemic role whatsoever in mathematical practice. What does a philosophical focus on *that* lead to?

My various expositions of the epistemic role puzzle were invariably accompanied by a *joke*. I've written, more than once: imagine that mathematical objects ceased to exist sometime in 1968. Mathematical work went on as usual. Why wouldn't it?[7] But one—I've since learned—should *never ever* make jokes in philosophical work. It's not merely that a dismayingly large number of professionals become totally lost when that happens; it's also that what's been written can (*will*) be taken literally. Perhaps it didn't help that some years later Balaguer wrote, echoing my earlier remarks, that: "[i]f all the objects in the mathematical realm suddenly *disappeared*, nothing would change in the physical world" (Balaguer 1998: 132).

Regardless. An interpretation of the epistemic role puzzle appeared in the literature that characterized it as a *modal* argument. The epistemic role puzzle was interpreted as trading on the construction of a possible world where there are no abstracta but the (empirical) world is otherwise exactly the same. This *makes no difference argument* has enjoyed a bit of discussion[8]; but the considerations both for and against "makes no difference" arguments are remote from what motivates the epistemic role puzzle.

First of all, I was offering a colorful thought experiment that was meant to draw the reader's attention to the considerations raised at the end of the last section: that epistemic access to the mathematical objects themselves (purportedly referred to by mathematical terms) plays no epistemic role in the practice of mathematics, neither (specifically) in how mathematical results are established, nor in (specifically) our studies of how mathematical practice operates. I wasn't suggesting the existence of a possible world in which there are no mathematical objects but mathematical practice goes on as before.[9] Baker writes of my thought experiment that:

> provoking the intuition that the existence of mathematical objects makes no difference by depending on thought-experiments whose conditions are conceptually impossible gives the [p]latonist plenty of ammunition for resistance.
>
> Baker (2010: 224)

Even if my aim were to ask the reader to imagine a possible world in which there are no mathematical objects, I hardly think that demands the reader to entertain conditions that are *conceptually impossible*. Perhaps this melodramatic phrasing is meant to allude to the common view that mathematical objects are metaphysically necessary. Even so, to imagine abstracta as not existing isn't to imagine something that's conceptually impossible. (Metaphysical necessity doesn't imply conceptual impossibility; surely we all know that by now.) Perhaps, instead, Baker means to suggest that imagining mathematical objects not to exist is to imagine certain mathematical truths (that are conceptually necessary) to instead be falsehoods. But this doesn't follow either. Our *understanding* of mathematical statements, and our understanding that such are true *doesn't* require that anything *exists*.[10] At least, this needs to be established by argument and not just assumed, as it seems to be by Baker.

## 1.3 McEvoy on the Epistemic Role Puzzle: His General Strategy

Claiming that sets, points and regions of spacetime are located within our perceptual ranges, I've noted, responds to Benacerraf's problem. That it doesn't respond to the epistemic role puzzle shows the puzzle to be distinct—according to whatever vague principles of individuation that *philosophical problems* seem to obey, anyway. One can't merely claim that points, regions of space, or sets *can* be perceived as Maddy and Field do in order to meet the epistemic role puzzle. One has to show that perception of these things actually *does* play a role in mathematical practice—in the proving of theorems, for example. (And, as we all also know by now: contexts of discovery don't count.) So "in principle" epistemic access to otherwise artificially-designed entities isn't *relevant*.

McEvoy, however, isn't out to show that the epistemic role puzzle reduces to Benacerraf's problem *tout court*; he only wants to show that the epistemic role puzzle raises no new considerations that *platonists* need worry about.[11]

McEvoy's argument for this claim is ingenious, intricate and detailed. Here's (roughly) how it goes. Consider the epistemic role puzzle all on its own. *Nothing* follows about the existence or non-existence of mathematical abstracta. After all, the epistemic role puzzle is only an *epistemic* observation, that there's no *epistemic* role for mathematical abstracta. So it's clear that claim alone can't imply that there are no mathematical abstracta, and *this* means that other premises are needed. McEvoy, by surveying the options that seem compatible with what I've written on this, argues that any other candidate premises—that would do the job to enable the epistemic role puzzle to refute the existence of mathematical entities—would manage that job *all on their own*. Therefore, argumentatively speaking, the epistemic role puzzle is an idle wheel.[12] McEvoy writes:

> The added premise, if ERP [Epistemic Role Puzzle] is to count as both a legitimate and novel challenge to platonism, must not be one the truth of which would itself refute platonism.
>
> McEvoy (2012: 298)

## 1.4 Correspondence Truth and Quine's Criterion, According to McEvoy

Unsurprisingly, the candidate additional premises that McEvoy considers as possible supplementations for the epistemic role problem involve metaphysical assumptions. Here's one:

> (AZ) The only non-epistemic role that abstracta can play is that of providing truth conditions for mathematical statements within a correspondence theory of truth, and the correspondence theory of truth is false.
>
> McEvoy op. cit.: 295

McEvoy writes:

> If AZ is added, it is no longer ERP that creates the problem for the platonist: it's the fact that platonism assumes the correspondence theory of truth, which, given AZ, is false. […] With the addition of AZ, though the argument from ERP does become valid, since AZ itself refutes platonism, this validity is purchased only at the cost of making ERP redundant.
>
> McEvoy op. cit.: 296

So too, consider a different premise:

> (AZ*) The only non-epistemic role that abstracta can play is that of entities over which the quantifiers of a suitably regimented best theory range, and this incurs a commitment to the untenable Quinean criterion of ontological commitment.
>
> McEvoy loc. cit.

McEvoy similarly observes:

> Once again the problem is that if AZ* is added, it is no longer ERP that creates the problem for the platonist: it is the assumption that the platonist must assume the (*ex hypothesi*) false Quinean criterion. Once again, when we add the premise to render the argument valid, ERP does none of the heavy lifting.
>
> McEvoy op. cit.: 297

What's an Azzouni to do? Well, here's what I suggest *this Azzouni* do: Deny that AZ and AZ* are to be interpreted as McEvoy does; in particular, McEvoy glosses the implications of rejecting the correspondence theory of truth and of rejecting Quine's criterion far too strongly. Appropriately glossed, *rejections* of these fundamental principles aren't sufficient on their own to refute platonism. What's needed to refute platonism is, along with the rejections of these principles, the epistemic role puzzle. I turn now to spelling out what I have in mind.

## 1.5 What Does Rejecting Correspondence Truth and Quine's Criterion Actually Imply?

Let's start with the rejection of the correspondence theory of truth (encapsulated in AZ) that, according to McEvoy, *all by itself* can be used to refute platonism. It should be admitted that the rejection of the correspondence theory of truth is *often* taken to be, as McEvoy seems to, the rejection of it with respect to *every* sentence of our language. Competing candidate theories of truth, when a philosopher is moodily fanatical this way, are also characterized globally. Coherence theories of truth, so interpreted—for example—deny that *any* sentence corresponds to

anything. Instead, *every* sentence (that's true) coheres—whatever that means exactly—with the other true sentences. So too, deflationary theories are interpreted by global truth fanatics as denying of *every* sentence that by describing it as a true one is ascribing a "substantial" truth-property to it. Deflationism, characterized this way, is militantly anti-metaphysical in intent.[13]

These alternatives to correspondence truth fit McEvoy's bill perfectly: if a platonist adopts one of *these* theories of truth, quick work can be made of a purported correspondence between mathematical truths and how it is with mathematical abstracta. But, of course, more *liberal* theories of truth are out there and flourishing. A candidate that's emerged recently is one or another species of pluralist truth[14]: Our language divides neatly into various discourses. Some operate according to correspondence principles of one or another sort, some operate according to coherence principles of one or another sort. Truth pluralists don't claim that platonism is ruled out just because correspondence truth has been ruled out for the language *as a whole*. Specific arguments about mathematical discourse are called for to adjudicate this question.

I'm not a fan of pluralism; but I *am* a proponent of a liberal version of deflationism.[15] *Liberal deflationism* treats the truth predicate as a logical device of blind ascription that's neutral with respect to whether true statements correspond to anything or not. "Sherlock Holmes is depicted as smarter than Mickey Mouse" and "Barack Obama is more famous than Saul Kripke" are both true; the second statement (in my view) corresponds to the way it is with several items in the world; this isn't true of the first statement.

This view of truth is neutral—as far as it goes—with respect to whether statements about abstracta are like statements about Mickey Mouse or like statements about Barack Obama (true because of correspondence to facts or true for some other reason). Liberal deflationism is a candidate theory of truth that's opposed to global correspondence; a proponent of it will accept AZ, but AZ doesn't, therefore, refute platonism. Additional premises are needed to manage this.

The same point can be made about Quine's criterion and McEvoy's AZ*, so I'll be brief. *Rejecting* Quine's criterion doesn't imply that quantifier statements[16] are *never* ontologically committing; it only implies that we can't assume that any particular quantifier statement *is* ontologically committing. Further premises, therefore, are needed as well to establish (or deny) that quantifier statements about abstracta are ontologically committing.

## 1.6 The Epistemic Role Puzzle to the Rescue

So what do the additional premises needed by the opponent of platonism look like? I imagine, of course, that there is more than one way to go here, but I want to sketch an argument that utilizes the epistemic role puzzle as the needed additional premise.

Start with the following *criterion for what exists*: anything that exists is mind- and language-independent. Next step: how do *we* recognize that an object is mind-

and language-independent? Answer: it has an *epistemic role*. Contrariwise, if an object has no epistemic role, then it's mind- and language-dependent, and therefore (by our criterion) it doesn't exist. Mathematical abstracta have no epistemic role. Conclusion: there are no mathematical abstracta.

I've defended detailed versions of this argument elsewhere,[17] and haven't the space to do it again here; I'll offer a couple of elucidations. First off, the criterion for what exists is one that I take us to have collectively adopted. It's not revealed by a conceptual analysis of ordinary words like "exist" or "there is" because the perceived meanings of those words are compatible with any number of criteria.[18]

Second, the argument as I've put it may seem to draw too strong a conclusion. Perhaps all that's licensed is the weaker conclusion that "we have *no reason to believe* in abstract objects, not that there are no such objects".[19] I think the stronger conclusion is licensed, however, because it's reasonable for us to say: *there are no Ss* if we have no reason to believe in *Ss*. Again, this is perhaps not the place to fully argue for this conclusion, but let me say this in my defense: it's sometimes reasonable for us to assert: *there are no Ss*. I think, for example, that it is reasonable for me to say: there is no Santa Claus, there are no hobbits, there is no Harry Potter. I think it's *understating* what I'm licensed to say if I say only: I've no reason to believe in these things. The latter is true, of course, and *it* is what licenses me to say that none of these things exist. Having said this, I should modestly add: I *could* be wrong. There might *be* a Santa Claus, a Harry Potter, some hobbits here and there. But surely the fact that I've no reason to believe in these things is compatible both with my being able to draw the conclusion: there are none of these things *and* I might be wrong about this. Surely the epistemic situation is the same with respect to *abstracta*.

## 1.7 Skepticism Again

This last point somewhat naturally brings us back to something I raised at the beginning of this paper, that the considerations behind the epistemic role puzzle (and behind Benaceraff's problem) aren't skeptical ones. McEvoy denies this, and as he notes, he has good company: Gödel, Katz, Burgess and Rosen (among others, no doubt). Giving his reasons for why the epistemic role puzzle as a challenge to the existence of platonic objects *is* analogous to skepticism, he writes (and I quote him nearly in full):

> The skeptical argument at play in this argument is that, for all we know, it is possible that there is, right now, no physical world, but that we are deluded into thinking there is one. This is a possibility since all of the evidence available in the event of the existence of the external world would be present in the skeptical world. The mathematician is similarly situated with respect to mathematical objects; given the lack of an epistemic role for mathematical objects, any evidence that would lead a mathematician to believe that there are mathematical entities would be present whether or not such entities exist. From this perspective, Azzouni's thought experiment begs the question against the argument from skeptical analogy: it assumes that the processes that produce our beliefs about the external world are currently operating reliably (i.e., they give us genuine knowledge of the external

> world). […] But the parallel between external world skepticism and mathematical skepticism that is relevant to the argument from skeptical analogy actually blocks this assumption. This parallel has to do with the possibility that we may *right now* be deceived about what is causing those experiences on the basis of which we infer the existence of physical reality, on the one hand, and mathematical reality on the other. Given this parallel, we cannot, as Azzouni does, hold constant our epistemic story of how these appearances are caused, and then ask what would happen if everything disappeared. This move is at once ruled out by the parallel that operates in the argument from skeptical analogy. We of course believe that we have knowledge of the external world, due to the reliable operation of our senses. (Similarly platonists believe that we have knowledge of the mathematical realm due to the reliable operation of intuition, or reason.) However, the point is that the falsity of this belief is compatible with our having a phenomenologically indistinguishable experience. This possibility of our having phenomenologically indistinguishable experiences regardless of whether entities of a certain kind exist is, after all, the point of skepticism; and the point of the argument from skeptical analogy is that this possibility obtains equally in the cases of mathematical and external world knowledge.
>
> McEvoy op. cit.: 302–303

How analogies between things may be drawn is, naturally enough, open-ended and fairly subjective (that's why analogies are so popular in *poetry*). Couple this point with my earlier hint that the individuation of philosophical problems is none too clear, and the reader won't be blamed for thinking that the ground rules of this debate between me and McEvoy are too ill-defined to be *resolvable*. Unsurprisingly, perhaps, I'm going to try to show this is false; that the matter isn't just resolvable, but resolvable in my favor.

Let's start with the background motivation for noting a parallel with traditional skepticism: it's a form of *dismissivism*. Correspondingly, the background motivation for denying the parallel to skepticism isn't a misguided hope for some credit for originality; it's the claim that even if skeptical considerations are *neutralized* (one way or another), the concern with the existence of mathematical abstracta *remains*. Notice that this neutralization point can hold even if there is an *analogy* between the two forms of argument. After all, surely there is an *analogy* (a rather close one) between what we might call *Evil-Demon skepticism* and *Dreaming skepticism*. Both involve possibility-scenarios (I might be dreaming right now; I might be in the grip of evil demon); nevertheless, it's obvious—or should be—that arguments that neutralize the threat of Dreaming skepticism might still leave Evil-Demon skepticism intact.

Let's start with Benacerraf's problem. The worry, as I've noted, is about a particular kind of *object*, an object that is causally inefficacious, causally insensitive, and not in space and time (and maybe that's weird in a whole bunch of other ways as well). It should be clear that even if one has *neutralized* the standard skeptical concerns McEvoy invokes (inferences from phenomenologically identical experiences that encapsulate all our relevant evidence), these concerns about these metaphysically peculiar objects remain. Imagine a scientist, for example, who postulates a kind of particle that is causally neutral in a similar way. It would be insensitive to the concerns of his opponents, *philosophically insensitive*, to invoke analogies to standard skepticism to undercut his concerns. To use faintly technical

terminology: this would be to *trivialize* those concerns. In short, Benacerraf's problem is rooted in the presumed metaphysical peculiarities of abstracta that make them immune to epistemic inroads. To compare this to skeptical scenarizing where one eliminates all our epistemic tools *altogether* (so that every object is now immune to epistemic inroads) misses the philosophical point.

Okay, what about the epistemic role puzzle? I would have hoped that the differences that matter between the concerns it raises and standard skeptical concerns would be even more obvious in its case. After all, the official concern of the puzzle is this: notice that our standard *epistemic practices* have certain accompaniments: methods of recognizing the *epistemic artifacts* that our means of access to the objects in question have *because of* those means of access. Indeed, the facts about this are intricate and subtle enough to give rise to *sciences of* those means of access. Surely, drawing an analogy between these considerations and external world skepticism is to miss the philosophical point: in the one case, we are focusing on facts about our epistemic access, facts that count *unfavorably* towards existence claims for mathematical abstracta. In the other case, we are undermining our methods of epistemic access altogether. The result, of course, counts unfavorably towards the existence of *anything* in the external world: this is *analogous*, in the usual sense of the word. I'd like to hope, however, that the *disanalogy* between undermining our methods of epistemic access altogether and noting that those very methods don't accommodate certain purported objects *even if left intact* would be taken note of as well.

## 1.8 Benacerraf's Problem and the Epistemic Role Puzzle: Close but Different Nevertheless

As I mentioned, the epistemic role puzzle can be seen as arising from Benacerraf's problem by flipping the focus of concern: not how do we manage to know about those objects, but instead, why isn't epistemic access to these objects a topic of mathematics, or a part of an ancillary science of mathematics? Nevertheless, in practice, bringing the puzzle against an approach in philosophy of mathematics can often be reconceptualized as bringing the problem against that approach. Doing so doesn't show they're the same concern; it only shows they live in the same neighborhood.

McEvoy writes of my discussion of apriorist varieties of platonism, when I complain that "no explanation is possible at all for how we can have a priori knowledge of ontologically independent objects …"[20]:

> In these passages, Azzouni raises the (serious) problem for the platonist's epistemology: how is it that human cognitive processes can reliably yield knowledge of a realm of abstracta which is entirely ontologically independent of those processes? If this sounds familiar, it is because we are now facing […] Field's version of the Benacerraf problem. Interpreting ERP as a demand for how our beliefs are sensitive to the facts obtaining in the

> platonic realm does yield a significant challenge to the [p]latonist […] but it does so at the cost of reducing ERP to the Benacerraf problem.
>
> McEvoy op. cit.: 299

I take myself to have *already established* in the foregoing pages of this paper that the puzzle and the problem are distinct, both in how they can be responded to, and in the sorts of issues that they make philosophically salient. My only job, therefore, is to delineate how these concerns dovetail in this particular case. The epistemic role puzzle tells us that there is no epistemic role for abstracta, in contrast with (some of) the items posited in the empirical sciences. It immediately follows that we have no *reliabilist story* for how we know what we (purportedly) know about abstracta *because we have no epistemic story at all*.[21] Benacerraf's problem runs the objection more directly: we have no reliabilist story for abstracta like we have in the empirical sciences. That's a *problem*.

## 1.9 Some Concluding Remarks

If you're a genuine (card-carrying) nominalist, Benacerraf's problem isn't going to quite do the job you need done. The contemporary proponents of Field's program show this by blithely countenancing abstracta, such as spatial points and regions—which they regard as meeting the challenge Benacerraf's problem successfully poses for more "remote" abstracta; so too, philosophers who think that mathematical statements can "index" nominalistic content—where such content involves spacetime-embodied abstracta — presumably feel that Benacerraf's problem is met by their abstracta, precisely because those items are spacetime-embodied.[22] The epistemic role puzzle is more unforgiving, so I've argued, requiring more of spatial points and regions—that they not only be "in principle" perceivable but that epistemic access to them actually play a role in the establishing of mathematical truths.[23]

Apart from this, however, I've argued that the epistemic role puzzle is a necessary component in at least one argument for nominalism because rejections of correspondence truth and Quine's criterion are insufficient all on their own to refute the platonist.

**Notes**

1. Benacerraf (1973).
2. See, e.g., Strobel and Heineman (1989) and Barret and Myers (2004).
3. See, e.g., Carey (2009), especially Chap. 4, for discussion of the literature.
4. See, e.g., the essays in Rock (1997).
5. See, e.g., Hacking (1983), Chap. 11.
6. See Azzouni (1994), Part I, § 5 and § 6 for details about these kinds of errors.
7. E.g., Azzouni (1994: 56, 2004a).

8. See Baker (2003) for the original coinage of the term "makes no difference," and for an attack on the argument. See Raley (2008) for a careful elucidation of the complex strands involved in these "MND" (Makes No Difference) arguments. She there labels an argument against the existence of mathematical abstracta that turns on calling the epistemic role puzzle the "epistemic version of MND"—I won't be using this terminology. I should add that I'm unsympathetic to the various MND arguments, largely for the reasons Raley gives for rejecting them. I won't discuss these details further now.
9. In Azzouni (1994), when I first offered this thought experiment, I was neutral on the question of the existence of abstracta—mainly because I already saw no reason to accept Quine's criterion. I first officially publicized my nominalism, however, in Azzouni (2004b). I thought the challenge posed by the epistemic role puzzle remained the same despite this change in viewpoint precisely because it wasn't (in my view) a philosophical claim but instead an (overlooked) insight about ordinary mathematical practice.
10. See Azzouni (2004b).
11. McEvoy (2012: 4), footnote 292.
12. This is almost his argument. Here's another rider he sometimes employs: where the epistemic role puzzle actually seems to do some work (in some versions of the argument against mathematical entities that McEvoy tries to mount on my behalf) it turns out that what it's doing is indistinguishable from the work Benacerraf's problem does.
13. See, e.g., Horwich (1998).
14. See Lynch (2004).
15. See Azzouni (2006, Forthcoming).
16. Statements of the form "$(\exists x)\ldots\ x\ \ldots$.".
17. E.g., in Azzouni (2004b), and in the appendix of the General Introduction in Azzouni (2010a).
18. See Azzouni (2007, 2010b).
19. McEvoy (2012: 291).
20. From Azzouni (2000: 237).
21. Furthermore—although I wasn't running this argument in 2000, see footnote 9—if our criterion for existence is mind- and language-independence, we can simply deny, on these grounds, that the purported objects exist.
22. See Daly and Langford (2009).
23. I should note, by-the-by, that I *don't* think that spacetime points or regions are "in principle" perceivable; but I've left that point aside for the sake of discussion.

## References


Azzouni, J. (1994). *Metaphysical myths, mathematical practice: The ontology and epistemology of the exact sciences*. Cambridge: Cambridge UP.

Azzouni, J. (2000). Stipulation, logic, and ontological independence. *Philosophia Mathematica*, *8* (3), 225–243.

Azzouni, J, (2004a). The derivation-indicator view of mathematical practice. *Philosophia Mathematica*, *12*(3), 81–105.

Azzouni, J. (2004b). *Deflating existential consequence: A case for nominalism*. Oxford: Oxford UP.

Azzouni, J. (2006). *Tracking reason: Proof, consequence, and truth*. Oxford: Oxford UP.

Azzouni, J. (2007). Ontological commitment in the vernacular. *Noûs*, *41*(2), 204–226.

Azzouni, J. (2010a). *Talking about nothing: Numbers, hallucinations, and fictions*. Oxford: Oxford UP.

Azzouni, J. (2010b). Ontology and the word 'exist': Uneasy relations. *Philosophia Mathematica*, *18*(3), 74–101.

Azzouni, J. (forthcoming). Deflationist truth. In (Michael Glanzberg, ed.) *Handbook on truth*. London: Blackwell.

Baker, A. (2003). Does the existence of mathematical objects make a difference? *Australasian Journal of Philosophy*, *81*(2), 246–264.

Baker, A. (2010). A medley of philosophy of mathematics: Review of philosophy of mathematics: Set theory, measuring theories, and nominalism. In G. Peter & G. Preyer (Eds.), *Metascience* (Vol. 19(2), pp. 221–224).

Balaguer, M. (1998). *Platonism and anti-platonism in mathematics*. Oxford: Oxford UP.

Barrett, H. H., & Myers, K. J. (2004). *Foundations of image science*. New York: Wiley.

Benacerraf, P. (1973). Mathematical truth. *The Journal of Philosophy 70* (19), 661–679.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford UP.

Daly, C., & Langford, S. (2009). Mathematical explanation and indispensability arguments. *The Philosophical Quarterly*, *59*(237), 641–658.

Field, H. H. (1989). *Realism, mathematics, and modality*. Oxford: Basil Blackwell.

Hacking, I. (1983). *Representing and intervening*. Cambridge: Cambridge UP.

Horwich, P. (1998). *Truth* (2nd Ed.). Oxford: Oxford UP.

Lynch, M. P. (2004). Truth and multiple realizability. *Australasian Journal of Philosophy*, *82*(3), 384–408.

Maddy, P. (1990). *Realism in mathematics*. Oxford: Oxford UP.

McEvoy, M. (2012). Platonism and the 'Epistemic Role Puzzle', *Philosophia Mathematica*, *20*(3), 289–304.

Raley, Y., (2008). Jobless objects: Mathematical posits in crisis. In G. Peters & G. Preyer (Eds.), *Philosophy of mathematics: Set theory, measuring theories, and nominalism* (pp. 112–131). Frankfurt: Ontos-Verlag.

Rock, I. (1997). *Indirect perception*. Cambridge, Mass: The MIT Press.

Strobel, H. A., & Heineman, W. R. (1989). *Chemical instrumentation: A systematic approach* (3rd Ed.). New York: Wiley.

# Chapter 2
# What Is the Benacerraf Problem?

**Justin Clarke-Doane**

In "Mathematical Truth," Paul Benacerraf presented an epistemological problem for mathematical realism. "[S]omething must be said to bridge the chasm, created by […] [a] realistic […] interpretation of mathematical propositions… and the human knower," he writes.[1] For *prima facie* "the connection between the truth conditions for the statements of [our mathematical theories] and […] the people who are supposed to have mathematical knowledge cannot be made out."[2]

The problem presented by Benacerraf—variously called "the Benacerraf Problem" the "Access Problem," the "Reliability Challenge," and the "Benacerraf-Field Challenge"—has largely shaped the philosophy of mathematics. Realist and antirealist views have been defined in reaction to it. But the influence of the Benacerraf Problem is not remotely limited to the philosophy of mathematics. The problem is now thought to arise in a host of other areas, including meta-philosophy. The following quotations are representative.

> The challenge for the moral realist […] is to explain how it would be anything more than chance if my moral beliefs were true, given that I do not interact with moral properties. […] [T]his problem is not specific to moral knowledge. […] Paul Benacerraf originally raised it as a problem about mathematics.
>
> Huemer (2005: 99)[3]

> It is a familiar objection to […] modal realism that if it were true, then it would not be possible to know any of the facts about what is […] possible […]. This epistemological

Thanks to Dan Baras, David James Barnett, Michael Bergmann, John Bigelow, John Bengson, Sinan Dogramaci, Hartry Field, Toby Handfield, Lloyd Humberstone, Colin Marshall, Josh May, Jennifer McDonald, Jim Pryor, Juha Saatsi, Josh Schechter, and to audiences at Australian National University, Monash University, La Trobe University, UCLA, the University of Melbourne, the University of Nottingham, the Institut d'Histoire et de Philosophie des Sciences et des Techniques and the University of Sydney for helpful comments.

J. Clarke-Doane (✉)
Columbia University, New York, USA
e-mail: justin.clarkedoane@gmail.com



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_2

> objection […] may […] parallel […] Benacerraf's dilemma about mathematical […] knowledge.
>
> Stalnaker (1996: 39–40)[4]

> We are reliable about logic. […] This is a striking fact about us, one that stands in need of explanation. But it is not at all clear how to explain it. […] This puzzle is akin to the well-known Benacerraf-Field problem […].
>
> Schechter: (2013: 1)[5]

> Benacerraf's argument, if cogent, establishes that knowledge of necessary truths is not possible.
>
> Casullo (2002: 97)

> The lack of […] an explanation [of our reliability] in the case of intuitions makes a number of people worry about relying on [philosophical] intuitions. (This really is just Benacerraf's worry about mathematical knowledge.)
>
> Bealer (1999: 52n22)

> [W]hat Benacerraf […] asserts about mathematical truth applies to any subject matter. The concept of truth, as it is explicated for any given subject matter, must fit into an overall account of knowledge in a way that makes it intelligible how we have the knowledge in that domain that we do have.
>
> Peacocke (1999: 1–2)[6]

One upshot of the discussion below is that even the above understates the case. An important class of influential but *prima facie* independent epistemological problems are, in relevant respects, restatements of the Benacerraf Problem. These include so-called "Evolutionary Debunking Arguments," associated with such authors as Richard Joyce and Sharon Street.

The Benacerraf Problem is, thus, of central importance. It threatens our knowledge across philosophically significant domains. But what exactly is the problem? In this paper, I argue that there is not a satisfying answer to this question. It is hard to see how there could be a problem that satisfies all of the constraints that have been placed on the Benacerraf Problem. If a condition on undermining, which I will call "Modal Security," is true, then there could not be such a problem. The obscurities surrounding the Benacerraf Problem infect all arguments with the structure of that problem aimed at realism about a domain meeting two conditions. Such arguments include Evolutionary Debunking Arguments. I conclude with some remarks on the relevance of the Benacerraf Problem to the Gettier Problem.

## 2.1 Benacerraf's Formulation

Benacerraf's "Mathematical Truth" has been deeply influential—but more for its theme than for its detail. The theme of the article is that there is a tension between the "standard" realist interpretation of mathematics and our claim to mathematical knowledge. Benacerraf writes,

> [O]n a realist (i.e., standard) account of mathematical truth our explanation of how we know the basic postulates must be suitably connected with how we interpret the referential apparatus of the theory. […] [But] what is missing is *precisely* […] an account of the link between our cognitive faculties and the objects known. […] We accept as knowledge only those beliefs which we can appropriately relate to our cognitive faculties.
>
> Benacerraf (1973: 674)

Benacerraf is skeptical that such an account exists. Thus, he thinks, we must either endorse a "non-standard" antirealist interpretation of mathematics or settle for an epistemic mystery.

What is Benacerraf's reason for being skeptical that our mathematical beliefs can be "appropriately related" to our cognitive faculties, if those beliefs are construed realistically? His reason is the causal theory of knowledge. He writes,

> I favor a causal account of knowledge on which for *X* to know that *S* is true requires some causal relation to obtain between *X* and the referents of the names, predicates, and quantifiers of *S*. [….] [But] […] combining *this* view of knowledge with the "standard" view of mathematical truth makes it difficult to see how mathematical knowledge is possible. […] [T]he connection between the truth conditions for the statements of number theory and any relevant events connected with the people who are supposed to have mathematical knowledge cannot be made out.
>
> Benacerraf (1973: 671–673)

There is a natural response to this argument. Even if the causal theory of knowledge were plausible in other cases, it seems inappropriate in the case of mathematics. As Øystein Linnebo remarks,

> By asking for a causal connection between the epistemic agent and the object of knowledge, Benacerraf treats […] mathematics […] like physics and the other […] empirical sciences. But […] [s]ince mathematics does not purport to discover contingent empirical truths, it deserves to be treated differently.
>
> Linnebo (2006: 546)

Indeed, not even the originator of the causal theory of knowledge, Alvin Goldman, intended that theory to apply to mathematics. In the article to which Benacerraf refers, Goldman begins,

> My concern will be with knowledge of empirical propositions only, since I think that the traditional [justified true belief] analysis is adequate for knowledge of non empirical truths.
>
> Goldman (1967: 357)

But the causal theory of knowledge is not even plausible in other cases. It does not seem to work with respect to knowledge of general truths or with respect to knowledge of truths about spatio-temporally distant events. Indeed, it has been almost universally rejected for reasons that are independent of the Benacerraf Problem (Goldman rejected the theory long ago).[7]

For these reasons, Benacerraf's own formulation of the problem no longer carries much weight. But it is widely agreed that Benacerraf was onto a genuine

epistemological problem for mathematical realism nevertheless. W. D. Hart summarizes the prevailing opinion nicely.

> [I]t is a crime against the intellect to try to mask the [Benacerraf] problem […] with philosophical razzle-dazzle. Superficial worries about […] causal theories of knowledge are irrelevant to and misleading from this problem, for the problem is not so much about causality as about the very possibility of natural knowledge of abstract objects.
>
> Hart (1977: 125–126)

The genuine problem to which Benacerraf was pointing is commonly thought to have been identified by Hartry Field. I turn to his formulation of the problem now.

## 2.2 Field's Improvement

Field's presentation of the Benacerraf Problem is the starting point for nearly all contemporary discussion of the issue[8]. It has a number of virtues which will occupy me below. Field writes,

> Benacerraf formulated the problem in such a way that it depended on a causal theory of knowledge. The [following] formulation does not depend on *any* theory of knowledge in the sense in which the causal theory is a theory of knowledge: that is, it does not depend on any assumption about necessary and sufficient conditions for knowledge.
>
> Field (1989: 232–233)

In particular,

> We start out by assuming the existence of mathematical entities that obey the standard mathematical theories; we grant also that there may be positive reasons for believing in those entities. These positive reasons might involve […] initial plausibility […] [or] that the postulation of these entities appears to be indispensable. […] But Benacerraf's challenge […] is to […] explain how our beliefs about these remote entities can so well reflect the facts about them […] *[I]f it appears in principle impossible to explain this*, then that tends to *undermine* the belief in mathematical entities, *despite* whatever reason we might have for believing in them.
>
> Field op. cit.: 26

Three observations about Field's formulation of the Benacerraf Problem are in order. First, as Field emphasizes, his formulation of the problem does not assume a view as to the necessary and sufficient conditions for knowledge. Field's formulation *does* assume a view as to the conditions that are (merely) necessary for *justification* (if justification is taken to be necessary for knowledge, then of course Field's formulation implies a necessary condition for knowledge too). According to Field, if one's beliefs from a domain $F$ are justified then it does not appear to her in principle impossible to explain the reliability of her $F$-beliefs.

Note that the claim is not—or, anyway, should not be—that if one's $F$-beliefs are justified, then one *can now explain* the reliability of one's $F$-beliefs. That would clearly be too stringent. Consider our perceptual beliefs. People's perceptual beliefs

were presumably justified before anything like an explanation of their reliability became available. Even today we have no more than a sketch of such an explanation. But it is less plausible that people's perceptual beliefs would have been justified if it appeared to them in principle impossible to explain the reliability of those beliefs.

The second observation is that Field's formulation of the Benacerraf Problem is *non-skeptical*. Field does not merely claim that it appears (to us realists)[9] in principle impossible to offer an explanation of the reliability of our mathematical beliefs *that would convince a mathematical skeptic*—one who doubts that there are any (non-vacuous) mathematical truths at all. Field *grants* for the sake of argument that our mathematical beliefs are both (actually) true and (defeasibly) justified, realistically conceived.[10] Field claims that, *even granted these assumptions*, it appears in principle impossible for us to explain the reliability of our mathematical beliefs.

This is important. In granting these assumptions, Field can draw a contrast between the likes of perceptual realism—realism about the objects of ordinary perception—and mathematical realism. Notoriously, it appears in principle impossible (for us realists) to offer an explanation of the reliability of our perceptual beliefs that would convince a perceptual skeptic. What we can arguably offer is an explanation of the reliability of our perceptual beliefs that assumes the reliability of our perceptual beliefs. We can arguably offer an evolutionary explanation of how we came to have reliable cognitive mechanisms for perceptual belief, and a neurophysical explanation of how those mechanisms work such that they are reliable.[11] Clearly, neither of these explanations would convince someone who was worried that we were brains in vats. The arguments for evolutionary theory and neurophysics blatantly presuppose the reliability of our perceptual beliefs. Still, these arguments do seem to afford our perceptual beliefs a kind of intellectual security. The question is whether analogous arguments are available in the mathematical case.[12]

The final observation is that, while Field does not seem to recognize it, his formulation of the Benacerraf Problem does not obviously depend on the view *that mathematics has a peculiar ontology*. *Prima facie*, his challenge merely depends on the view that mathematical truths are causally, counterfactually, and constitutively independent of human minds and languages.[13] The converging opinion that there is no epistemological gain to "trading" ontology for ideology in the philosophy of mathematics reflects this point. But the point is often misconstrued.[14] The point is not that the explication of the ideological "primitives" will still somehow make reference to abstract objects, so the apparent loss of ontology is illusory. The point is that abstract objects are not what give rise to the Benacerraf Problem in the first place.

This point should not be surprising. As was mentioned in the introduction, something similar to the Benacerraf Problem is commonly thought to arise for realism about domains like morality, modality, and logic. But none of these domains, at least obviously, has a peculiar ontology. In the context of nominalism about universals, morality merely carries with it additional ideology ("is good," "is bad," "is obligatory," and so on). Similarly, for one who takes modal operators as

primitive, the same is true of modality. Finally, it is certainly not mandatory to think that there are peculiar objects corresponding to (first order) logical (as opposed to metalogical) truths. However, in none of these cases does the existence of something like a Benacerraf Problem seem to depend on the plausibility of an ontologically innocent interpretation of the corresponding domain.[15]

The canonical formulation of the Benacerraf Problem, due to Field, is, thus, appealing. It does not rely on a theory of knowledge, much less a causal one. It does not simply raise a general skeptical problem for mathematical realism that has an analog in the perceptual case. Finally, it does not, in any obvious way, rely on an ontologically committal interpretation of mathematics. Nevertheless, it is unclear at a crucial juncture. It is unclear what it would take to *explain the reliability* of our mathematical beliefs in the relevant sense. In what sense of "explain the reliability" is it plausible *both* that it appears in principle impossible (for us realists) to explain the reliability of our mathematical beliefs, *and* that the apparent in principle impossibility of explaining the reliability of those beliefs undermines them?

## 2.3 Safety

In addressing this question, it will be helpful to begin by considering an account of mathematical truth that even Field believes meets his challenge. We can then look for the sense in which this view can "explain the reliability" of our mathematical beliefs and in which the apparent in principle impossibility of explaining their reliability would undermine them.

The view in question is a version of mathematical pluralism. The key idea to this view is that consistency suffices for truth in mathematics. This contrasts with "standard" mathematical realism, according to which the overwhelming majority of consistent (foundational) mathematical theories are false (just as the overwhelming majority of consistent physical theories are false). Although this view takes many forms [see Carnap [1950] (1983); Putnam [1980] (1983); Linsky and Zalta (1995); and Hamkins (2012)], Field has been clearest about the merits of Mark Balaguer's "Full Blooded Platonism" (FBP). According to FBP, consistent mathematical theories are automatically about the class of objects of which they are true, and there is always such a class (where consistency is a primitive notion, and the notion of truth is a standard Tarskian one).[16]

FBP was specifically advanced as a solution to Field's formulation of the Benacerraf Problem. Mark Balaguer writes,

> The most important advantage that FBP has over non-full-blooded versions of platonism […] is that all of the latter fall prey to [Field's formulation of] Benacerraf's epistemological argument.
>
> Balaguer (1995: 317)

Balaguer explains,

> [FBP] eliminates the mystery of how human beings could attain knowledge of mathematical objects. For if FBP is correct, then all we have to do in order to attain such knowledge is conceptualize, or think about, or even "dream up," a mathematical object. Whatever we come up with, so long as it is consistent, we will have formed an accurate representation of some mathematical object, because, according to FBP, all [logically] possible mathematical objects exist.
>
> Balaguer loc. cit.

Field agrees. He writes,

> [Some philosophers] (Balaguer (1995); Putnam [1980] (1983); perhaps Carnap [1950] (1983) solve the [Benacerraf] problem by articulating views on which though mathematical objects are mind-independent, any view we had had of them would have been correct […]. [T]hese views allow for […] knowledge in mathematics, and unlike more standard Platonist views, they seem to give an intelligible explanation of it.
>
> Field (2005: 78)

In what relevant sense of "explain the reliability" does FBP explain the reliability of our mathematical beliefs? The quotations above suggest that FBP explains the reliability of our mathematical beliefs in the sense that it shows *that had our mathematical beliefs been different (but still consistent), they still would have been true*. However, FBP only shows this assuming that the mathematical truths are the same in the nearest worlds in which our mathematical beliefs are different.[17] If there are no (existentially quantified) mathematical truths in the nearest worlds in which we have different mathematical beliefs, for instance, then had our mathematical beliefs been (appropriately) different, they would have been false. I shall assume, then, that Field intends to grant not just the (actual) truth of our mathematical beliefs, but also that the mathematical truths are the same in all nearby worlds.

Would an explanation in the relevant sense need to show that had our mathematical beliefs been different, they still would have been true? Surely not. Consider the perceptual case. Had our perceptual beliefs been (sufficiently) different (but still consistent), they would have been false. The closest worlds in which we have (consistent) perceptual beliefs as of goblins, say, is a world in which we are deluded somehow. Perhaps it is true that had our perceptual beliefs been different, but "similar," they still would have been true (assuming that there is some independent way to explicate the notion of similarity). But it is still doubtful that an explanation in the relevant sense would need to show this. The observation that had our beliefs of a kind $F$ been different, they would have been false only seems undermining *to the extent that they could have easily been different*. If the closest worlds in which our $F$-beliefs are different but "similar" are remote, then it is hard to see how the observation that had we been in those worlds, our $F$-beliefs would have been false, could undermine them.

The reasonable explanatory demand in the neighborhood, which FBP does seem to address (if the mathematical truths are the same in all nearby worlds), is to show that our mathematical beliefs are *safe*—i.e. to show that we could not have *easily*

had false mathematical beliefs (using the method we used to form ours).[18] For typical contingent truths $F$, our $F$-beliefs can fail to be safe in two ways (assuming their actual truth). They can fail to be safe, first, if it could have easily happened that the *F-truths* were different while our $F$-beliefs failed to be correspondingly different. They can fail to be safe, second, if it could have easily happened that our *F-beliefs* were different while the $F$-truths failed to be correspondingly different. If, however, the $F$-truths could not have easily been different, then our $F$-beliefs cannot fail to be safe in the first way. If, moreover, the $F$-truths are "full-blooded," in the sense that every (consistent) $F$-theory is equally true, then our $F$-beliefs cannot fail to be safe in the second way (assuming the "safety" of our inferential practices). Thus, if the mathematical truths are the same in all nearby worlds, and we are granted the (actual) truth of our mathematical beliefs, then FBP shows that our mathematical beliefs are safe.

But if the mathematical truths are the same in all nearby worlds, and we are granted the (actual) truth of our mathematical beliefs, then our mathematical beliefs may well be safe *even if standard mathematical realism is true*. Again, if the mathematical truths could not have easily been different (whether or not they are full-blooded), then our mathematical beliefs cannot fail to be safe in the first way (assuming their actual truth). Moreover, *if we could not have easily had different mathematical beliefs* (even if, had we, they would not still have been true), then they cannot fail to be safe in the second way (again, assuming their actual truth). But there are reasons to think that we could not have easily had different mathematical beliefs. Our "core" mathematical beliefs might be thought to be evolutionarily inevitable.[19] Given that our mathematical theories best systematize those beliefs, there is a "bootstrapping" argument for the safety of our belief in those theories. Our "core" mathematical beliefs are safe; our mathematical theories "abductively follow" from those; our abductive practices are "safe" (something Field, as a scientific realist, would presumably concede); so, our belief in our mathematical theories is safe.

Of course, to the extent that our mathematical theories do not best systematize our "core" mathematical beliefs, this argument is not compelling. But, then, to that extent, the standard realist should not believe in those theories anyway. The reason that standard realists typically refuse to endorse either the Continuum Hypothesis (CH) or its negation is precisely that neither CH nor not-CH seems to figure into the (uniquely) best systematization of our mathematical beliefs.

This argument for the safety of our mathematical beliefs obviously turns on speculative empirical hypotheses. In particular, what evolutionary considerations most clearly suggest is that we would have certain quantificational and geometrical beliefs about things in our environments in certain situations. It does not seem that it was evolutionarily inevitable for us to have the "core" pure mathematical beliefs that we have. One way to address this problem would be to argue that our pure mathematical theories plus bridge laws linking pure mathematical truths to quantificational and geometric truths about things in our environments "abductively follow" from the latter. But the point is that there is a promising argument here—one that it does not appear "in principle impossible" to make.

It appears, then, that the standard mathematical realist may also be able to show that our mathematical beliefs are safe. Is there some *other* relevant sense of "explain the reliability" in which FBP can explain the reliability of our mathematical beliefs? I cannot think of one. (If there is not, then there may be little to recommend FBP.) But there is another epistemological challenge which is closely related to the challenge to show that our mathematical beliefs are safe. Let me turn to that now.

## 2.4 Sensitivity

Despite his remarks on mathematical pluralism, Field typically suggests that his challenge is to show that our mathematical beliefs are *counterfactually persistent*—i.e., that *had the mathematical truths been (arbitrarily) different (or had there been no existentially quantified such truths at all), our mathematical beliefs would have been correspondingly different* (if we still formed our mathematical beliefs using the method we actually used to form them). For example, Field writes,

> The Benacerraf problem […] seems to arise from the thought that we would have had exactly the same mathematical […] beliefs even if the mathematical […] truths were different […] and this undermines those beliefs.
>
> Field (2005: 81)[20]

Notice that FBP does nothing to answer this challenge. Balaguer is "doubtful that mathematical theories are necessary in any interesting sense" (Balaguer 1998: 317), and he concedes that "[i]f there were never any such things as [mathematical] objects, the physical world would be exactly as it is right now" (Balaguer 1999: 113). Given the supervenience of the intentional on the physical, it follows that had there been no existentially quantified mathematical truths, our mathematical beliefs would have failed to be correspondingly different.

Would the apparent in principle impossibility of showing that our mathematical beliefs are counterfactually persistent in this sense undermine them? Surely not. We cannot even show that our perceptual beliefs are counterfactually persistent in *this* sense. As skeptics argue, the closest worlds in which the perceptual truths are sufficiently different is a world in which we are deluded.[21]

The reasonable challenge in the neighborhood is to show that our mathematical beliefs are *sensitive*—i.e., that *had the contents of our mathematical beliefs been false, we would not still have believed them* (using the method we actually used to form our mathematical beliefs). Although we cannot show that had the perceptual truths been (arbitrarily) different, our perceptual beliefs would have been correspondingly different, Nozick (1981) pointed out that we can, it seems, show that had I, e.g., not been writing this paper, I would not have believed that I was. The closest world in which I am not writing this paper is still a world in which my perceptual faculties deliver true beliefs.

But there is a problem with Field's challenge under this construal. Contra Balaguer, mathematical truths seem to be metaphysically necessary. If they are,

then our mathematical beliefs are vacuously sensitive on a standard semantics. As David Lewis writes,

> [I]f it is a necessary truth that so-and-so, then believing that so-and-so is an infallible method of being right. If what I believe is a necessary truth, then there is no possibility of being wrong. That is so whatever the subject matter […] and no matter how it came to be believed.
>
> Lewis (1986: 114–115)[22]

Field might respond that, though we are granted the *actual* truth (and defeasible justification) of our mathematical beliefs (and presumably also that the mathematical truths could not have easily been different), we are not granted the necessity of their contents. Unlike the claim that the contents of our mathematical beliefs are true, the claim that their contents are necessary requires argument. But our belief that the mathematical truths are necessary is commonly thought to enjoy a similar status to our belief that they are true. Both beliefs are commonly regarded as default-justified positions. If Field were merely trying to undermine our mathematical beliefs *under the assumption that our belief that they are necessary is not itself (defeasibly) justified*, then the interest of his challenge would be greatly diminished.[23]

Of course, it is arguable, contra Lewis, that some conditionals with metaphysically impossible antecedents are not vacuously true. For a variety of purportedly metaphysically necessary truths, we seem to be able to intelligibly ask what would have been the case had they been false. As Field writes,

> If one [says] "nothing sensible can be said about how things would be different if the axiom of choice were false," it seems wrong: if the axiom of choice were false, the cardinals wouldn't be linearly ordered, the Banach-Tarski theorem would fail and so forth.
>
> Field (1989: 237–238)

However, we seem to be equally unable to show that our relevantly uncontroversial beliefs are non-vacuously sensitive. For example, we seem to be unable to show that our belief in "bridge laws" that link supervenient properties to subvenient ones are so sensitive. Had—as a matter of metaphysical impossibility—atoms arranged car-wise failed to compose cars (as "ontological nihilists" allege), we still would have believed that they did.[24]

It might be thought that Field could simply accept that his challenge is at least as serious for realism about truths that link supervenient properties to subvenient ones. After all, such truths are metaphysically necessary, and, again, the Benacerraf Problem is often thought to arise for realism about such truths. The problem with this response is that it would not just require rejecting realism about necessary truths. It would at least *prima facie* require rejecting realism about the truths of ordinary perception. If our belief about the composition conditions of cars is undermined, then so too, presumably, is our belief that we are not sitting in one.[25] If Field's challenge generalized this wildly, then it would no longer point to an epistemic difference between our mathematical beliefs and our beliefs about the objects of ordinary perception.

I will shortly mention a way to argue that our mathematical beliefs are even *better* off than our beliefs about the composition conditions of cars with respect to sensitivity. But the above demonstrates that if Field's challenge is to show that our mathematical beliefs are sensitive, then, again, it either does not appear in principle impossible to answer, or this appearance does not plausibly undermine those beliefs.

## 2.5 Indispensability

Those familiar with "Evolutionary Debunking Arguments" (to be discussed) are likely to think that I have failed to consider the most obvious analysis of Field's challenge. The challenge, it might be thought, has nothing immediately to do with the safety and sensitivity of our mathematical beliefs. It has to do with the explanation of our having those beliefs. The challenge is to show *that the contents (or truth) of our mathematical beliefs figure into the best explanation of our having them*.

This proposal is related to a causal constraint on knowledge. But it is intuitively weaker. Rather than requiring that the subject matter of our beliefs from an area *F* helps to *cause* our having the *F*-beliefs that we have, it is required that the contents of our *F*-beliefs help to *explain* our having those beliefs. *Prima facie*, our beliefs may have a causally inert subject matter, though their contents figure into the best explanation of contingent states like our having those beliefs.

But there is an obvious problem with Field's challenge under this analysis. It simply does not appear in principle impossible to answer. Insofar as mathematics appears to be indispensable to empirical science, it appears impossible to show that the contents of our mathematical beliefs do *not* figure into the best explanation of our having them. As Mark Steiner writes,

> [S]uppose that we believe […] the axioms of analysis or of number theory. […] [S]omething is causally responsible for our belief, and there exists a theory — actual or possible, known or unknown — which can satisfactorily explain our belief in causal style. This theory, like all others, *will contain the axioms of number theory and analysis*.
>
> Steiner (1973: 61)

The point is that mathematics, like logic, seems to be assumed by all of our empirical theories.[26] If it is, however, then mathematics is a background assumption of the theory that best explains our having the mathematical beliefs that we have. In particular, for any typical (e.g. not higher-set-theoretic) mathematical proposition *p*, *p* is a background assumption of the best explanation of our having the belief that *p*.

One might respond to this problem by arguing that an explanation in the pertinent sense would show, not just that the contents of our mathematical beliefs figure into the best explanation of our having those beliefs, but that they do so in an "explanatory way." But, setting aside the obscurity of the quoted locution, it would still not appear in principle impossible to explain the reliability of our mathematical

beliefs. The key point of the "improved" indispensability argument pressed lately by Alan Baker and others is that mathematical hypotheses seem to figure into the best explanation of empirical phenomena in just such a way.[27] This argument does not show that the contents of our mathematical beliefs figure into the best explanation *of our having them* in an explanatory way. But the latter claim is *prima facie* plausible. Consider, for instance, Sinnott-Armstrong's suggestion that "[p]eople evolved to believe that 2 + 3 = 5, because they would not have survived if they had believed that 2 + 3 = 4" (Sinnot-Armstrong 2006: 46). The question arises: why would people not have survived if they had believed that 2 + 3 = 4? According to Sinnott-Armstrong "the reason why they would not have survived then is that it is true that 2 + 3 = 5" (Sinnot-Armstrong loc. cit.). Sinnott-Armstrong appears to be suggesting that 2 + 3 = 5 explains our ancestors' coming to believe that 2 + 3 = 5. It does not merely "figure into" the explanation of their doing this. Whether this is true is certainly debatable. The point, however, is that it does not appear "in principle impossible" to show that the contents of our mathematical beliefs figure into the best explanation of our having them—even "in an explanatory way."[28]

Field is under no illusions on this point. When discussing his challenge as it applies to logic, he explicitly rejects the above analysis of his challenge (Field 1996: 372n13). Field even notes that the realist can appeal to the explanatory role of mathematics in an effort to bolster the conclusion of the previous section. He notes, in effect, that one can argue from explanatory considerations to the *non-vacuous* sensitivity of our mathematical beliefs. Field writes,

> [O]ne can try to invoke indispensability considerations […] in the context of explaining reliability. One could argue […] that if mathematics is indispensable to the laws of empirical science, then *if the mathematical facts were different, different empirical consequences could be derived from the same laws of (mathematicized) physics*.
>
> Field (1989: 28)

Such an argument for the sensitivity of our beliefs about the composition conditions of ordinary objects would not be plausible. When $p$ is a truth predicating a property—such as the property of being a car—which is not the "postulate" of any special science, one cannot argue from the explanatory indispensability of $p$ to the sensitivity of our belief that $p$, as above. This is true even if the property is supervenient on properties which are the "postulates" of special sciences.[29]

What should we think of the above argument? I need offer no final assessment here. But Field's reason for doubt is not compelling. He objects that "the amount of mathematics that gets applied in empirical science… is relatively small" (Field 1989: 29). But, as with safety, there is a "bootstrapping" argument from the sensitivity of our belief in applicable mathematics to the sensitivity of our belief in the rest of it. Again, such an argument may fail to "decide" certain abstract hypotheses. But this would merely seem to confirm the standard view that we ought to remain agnostic about their truth-values.

Suppose, however, that it *did* appear in principle impossible to show that the contents of our mathematical beliefs figure into the best explanation of our having

them. Would this appearance undermine those beliefs? It is hard to see how it could. I have argued that we may be able to show that our mathematical beliefs are both safe and sensitive, given their (actual) truth (and defeasible justification), in the sense in which we can show that our uncontroversial beliefs are. We may be able to argue that we could not have easily had false mathematical beliefs, given their truth (and that the mathematical truths could not have easily been different), because we could not have easily had different ones. Thus, our mathematical beliefs are safe. And we may be able to argue that, since mathematical truths would be metaphysically necessary, had the contents of our mathematical beliefs been false, we would not have believed those contents (in the sense that we can argue that, had the contents of our explanatorily basic ordinary object beliefs been false, we would not have believed those contents). Thus, our mathematical beliefs are (vacuously) sensitive as well. Notice that this argument does not assume, even implicitly, that the contents of our mathematical beliefs figure into the best explanation of our having them. The problem, then, is this. How could the observation that the contents of our mathematical beliefs fail to figure into the best explanation of our having them undermine them *without giving us some reason to doubt that they are both safe and sensitive*? This obscurity points to a basic problem with the Benacerraf Problem.

## 2.6 In Search of a Problem

What is the Benacerraf Problem? It is not to show that our mathematical beliefs are safe or sensitive, since we may be able to show that they are safe and sensitive in the sense that we can show that our uncontroversial beliefs are. Nor is it to show that the contents of our mathematical beliefs figure into the best explanation of our having them, since we may be able to show that too—and, anyway, it is unclear how the apparent in principle impossibility of showing this could undermine our mathematical beliefs. Is there some *other* sense of "explain the reliability" in which it appears in principle impossible to explain the reliability of our mathematical beliefs and which is such that this appearance plausibly undermines those beliefs?

It is sometimes said that what is wanted is a *unified* explanation of the correlation between our mathematical beliefs and the truths. Without this, that correlation seems like a "massive coincidence."[30] But what does this worry amount to?

It cannot be that the reliability of our mathematical beliefs would be *improbable*. Either the probability at issue is epistemic or it is "objective." If it is epistemic, then the suggestion is question-begging. It effectively amounts to the conclusion that Field's argument is supposed to establish—that our mathematical beliefs are not justified. Suppose, then, that the probability is objective. Then for any mathematical truth $p$, presumably $\Pr(p) = 1$, given that such truths would be metaphysically necessary.[31] Moreover, as we have seen, it may be that Pr (we believe that $p$) $\approx 1$, because the probability of our having the mathematical beliefs that we have is high.[32] But then, Pr ($p$ & we believe that $p$) $\approx 1$, by the probability calculus. Since

($p$ & we believe that $p$) implies that (our belief that $p$ is true), it may be true that Pr(our belief that $p$ is true) $\approx 1$.

Nor can the request for a unified explanation be the request for a *common cause* of the mathematical truths and of our mathematical beliefs. Such a request blatantly assumes a causal constraint on knowledge, which Field's challenge is supposed to avoid. (Nor, again, can the request for a unified explanation be a request to show that the contents of our mathematical beliefs help to explain—even if not to cause—our having them. Again, the contents of our mathematical beliefs may well do that.)[33]

The problem is not just that it is hard to state an analysis of Field's challenge that satisfies the constraints that he places on it. It is not clear that there *could be* a problem that satisfies those constraints, given what I argued in Sects. 2.3 and 2.4.

To see this, let me introduce a condition on information $E$ if it is to undermine our beliefs of a kind $F$. The condition is:

> *Modal Security*: If information $E$ undermines all of our beliefs of a kind $F$, then it does so by giving us reason to doubt that our $F$-beliefs are both sensitive and safe.[34]

Modal Security states a *necessary* condition on underminers. It does not say that if information $E$ gives us reason to doubt that our $F$-beliefs are both safe and sensitive, then $E$ obligates us to give up all of those beliefs. It says that if $E$ does not even do this, then $E$ cannot be thought to so obligate us.[35]

The key idea behind Modal Security is that there is no such thing as a "non-modal underminer." (Of course, there is such a thing as a non-modal defeater—namely, a rebutter, i.e. a "direct" reason to believe the negation of the content of our defeated belief.[36]) If there were such a thing as a non-modal underminer, then information could undermine our beliefs "immediately." It could undermine them, but not by giving us reason to doubt that they are modally secure. The "by" is needed, since everyone should agree that if $E$ undermines our $F$-beliefs, then $E$ gives us reason to doubt that those beliefs are safe—i.e. that they could not have easily been false. If our $F$-beliefs are actually false, then they could have easily been.

Paradigmatic underminers seem to conform to Modal Security. If $E$ is that we took a pill that gives rise to random $F$-beliefs, for instance, then $E$ seems to undermine them by giving us reason to believe that we could have easily had different $F$-beliefs, and, hence (given that the $F$-truths do not counterfactually depend on our $F$-beliefs), that our $F$-beliefs are not safe. Something similar could perhaps be said of evidence $E$ that there is widespread disagreement on $F$-matters (though whether such evidence is undermining is of course debatable). On the other hand, when the $F$-truths are contingent, and $E$ is that we were "bound" to have the $F$-beliefs that we do have because of some constraining influence, such as a tendency to overrate one's self, then $E$ seems to undermine our $F$-beliefs by giving us reason to doubt that had the contents of our $F$-beliefs been false, we would not still have believed those contents—i.e., by giving us reason to doubt that our $F$-beliefs are sensitive.

Nevertheless, it might be thought that Modal Security cannot handle necessary truths that we were "bound" to believe. Suppose that a machine enumerates sentences, deeming them validities or invalidities. Independent investigation has confirmed its outputs prior to the last five. The last five outputs are "validity." We defer to the machine's last five outputs only, and have no prior metalogical beliefs. Today a trusted source tell us that the machine was "stuck" in the last five instances. Call this evidence *E*. Then *E* does not seem to give us "rebutting" reason to believe that the last five outputs are invalidities. Nor does it seem to give us undermining reason to doubt that had, as a matter of metaphysical possibility, the last five outputs—those sentences—been invalidities, the machine would not have called them validities. (*E* does not seem to give us reason to believe that there is a metaphysically possible world in which those sentences are invalidities, and the counterfactual in question can only be false with respect to such worlds if there is.) Nor does *E* seem to give us undermining reason to believe that the machine could have easily called the last five sentences invalidities. That was the point of calling it "stuck." But *E* seems to undermine all of our metalogical beliefs. Does not this show that *E* may undermine all of our beliefs of a kind but not by giving us reason to doubt that they are both sensitive and safe?[37]

It does not. *E* may not be evidence that, for any metalogical proposition *p that we believe*, we could have easily had a false belief as to whether *p*. It does not follow that *E* is not evidence that we could have easily had a false metalogical belief. *E* is evidence that, even if the machine had considered an invalidity last, it still would have called the sentence a validity. But it is not just this fact which seems undermining. If we know that the only worlds in which it considers an invalidity last are "distant," then it is hard to see how evidence that, had we been in one, we would have had false metalogical beliefs, could undermine all of our metalogical beliefs. It must apparently be added that we know that such worlds are "near" to the actual one.[38] But if this is added, then *E* is evidence that we could have easily had different metalogical beliefs, had the machine considered different sentences last, and hence, given the necessity of the metalogical truths, that we could have easily had false metalogical beliefs—i.e., that those beliefs are not safe.[39]

This response depends on the assumption that not just any grouping of beliefs counts as a "kind." If we could let *F* be $\{[b] = \{x: x = b\}\}$, for some belief formed by the machine, *b*, then *E* could undermine all of our *F*-beliefs despite giving us no reason to believe that we might have easily had false *F*-beliefs. Intuitively, however, metalogical beliefs, like moral beliefs, are kinds, while "the last metalogical belief formed by the machine" is not. If there is no principled argument for this, however, then Modal Security may not get off the ground. The problem is similar to the "generality problem" for process reliabilism.[40]

Of course, if we could explicate "stuck" in such a way that learning that the machine was stuck in the last five instances undermines all of our metalogical beliefs, but has no modal implications at all, then the example above would be a

counterexample to Modal Security. But it is hard to see how we could do this. We might say that the machine is "stuck" in that it is not "detecting," "tracking," or "sensitive to" the metalogical truths—it is not generating its outputs "because" they are true. But what do these locutions mean? They do not mean that the truth of the machine's outputs is not implied by their best explanation. Had we imagined instead a machine that outputs only logical truths themselves, then, trivially, the machine would output such truths "because" they were true, since every logical truth is a consequence of every explanation at all. We might explicate "stuck" in terms of hyperintensional ideology like constitution or ground. But why exactly should give up beliefs which we learn are not "constituted by" or "grounded in" their truth?[41]

The relevance of Modal Security to the Benacerraf Problem is as follows. If it is true, and what was said in Sects. 2.3 and 2.4 is correct, then there could not be a problem that plays the epistemological role that the Benacerraf problem is supposed by Field to play. Even if there is a sense in which it appears in principle impossible to "explain the reliability" of our mathematical beliefs, this is not a sense which gives us reason to doubt that they are both safe and sensitive (with respect to metaphysically possible worlds). Field does not even pretend to give us ("rebutting") reason to think that they are actually false, and he does not seem to give us any ("undermining") reason to doubt that if they are true, then they are both safe and sensitive. Hence, even if there is a sense in which it appears in principle impossible to "explain the reliability" of our mathematical beliefs, this is not a sense which undermines them, if Modal Security is true.

To be sure, the conclusions of Sects. 2.3 and 2.4 are subject to amendment. As more is learned about the genealogy of our mathematical capacities, it may come to appear impossible to show that our "core" mathematical beliefs are inevitable. But if we can show that they are, and if Modal Security is true, then we can "explain their reliability" in every sense which is such that the apparent in principle impossibility of explaining their reliability undermines them.

Note that this is *not* to say that our mathematical beliefs are in good epistemic standing—anymore than it is to say that the theological beliefs of a theological realist who can argue both that the theological truths would be metaphysically necessary if true, and that she could not have easily had different theological beliefs, are in good epistemic standing. For all that has been said, our mathematical beliefs could be false and unjustified. Field *grants* for the sake of argument that our mathematical beliefs are (actually) true and (defeasibly) justified, in order to generate a *dialectically effective* argument against realists. But if Modal Security is true, then such an argument may not, in general, be possible. Modal Security implies that Field overreaches.

Let me illustrate the above reasoning with reference to two further analyses of Field's challenge. When discussing the Lewisian reply to the challenge to show that our mathematical beliefs are sensitive, Field proposes an alternative. He writes,

> If the intelligibility of talk of "varying the facts" is challenged […] it can easily be dropped without much loss to the problem: there is still the problem of explaining the *actual* correlation between our believing "*p*" and its being the case that *p*.
>
> Field (1989: 238)

The problem is this. Even if there is some hyperintensional sense of "explanation" according to which one can intelligibly request an explanation of the "merely actual correlation" between our mathematical beliefs and the truths, it is unclear how the apparent in principle impossibility of offering that could undermine those beliefs—given that we may still be able to show that they are safe and sensitive. If we can show that our mathematical beliefs are safe and sensitive, given their truth, then we can show that they were (all but) *bound* to be true.

Schechter suggests another response to Lewis on behalf of Field. He writes,

> Lewis is correct […] that the reliability challenge for mathematics […] is subject to a straightforward response, so long as the challenge is construed to be that of answering [the question of how our mechanism for mathematical belief works such that it is reliable] […]. [But] [t]here remains the challenge of answering [the question of how we came to have a reliable mechanism for mathematical belief].
>
> Schechter (2010: 445)

Schechter claims that the question of *how we came to have* a reliable mechanism for mathematical belief may remain open even under the assumption that it is unintelligible to imagine the mathematical truths being different (even if the question of *how that mechanism works* such that it is reliable may not). But, first, this appears incorrect. Schechter is explicit that the question of how we came to have a reliable mechanism for mathematical belief is different from the question of how we came to have the mechanism for mathematical belief *that we actually came to have* (since the latter question is clearly answerable in principle). However, in order to decide whether we were, say, selected to have a reliable mechanism for mathematical belief, as opposed to being selected to have a mechanism for mathematical belief with property *F which is in fact reliable*, we would seem to need to have to decide what mechanism it would have benefited our ancestors to have had *had the mathematical truths been different*.[42] Second, even if Schechter were correct, it is hard to see how the apparent in principle impossibility of explaining the reliability of our mathematical beliefs in *his* sense could undermine them.[43] For all that has been said, we might still be able to show that our mathematical beliefs are safe and sensitive.

Whether Modal Security is true requires in-depth treatment. But while successfully defending it would suffice to deflate the Benacerraf Problem, successfully challenging it would not suffice to reestablish that problem. The question would remain: *in what sense of "explain the reliability" is it plausible both that it appears in principle impossible to explain the reliability of our mathematical beliefs and that the apparent in principle impossibility of explaining their reliability undermines them?* Insofar as there seems to be no satisfactory answer to this question, the force of the Benacerraf Problem seems lost. Of course, it does not follow that there

are no mysteries surrounding mathematical knowledge. The point is that no such mystery can play the role that the Benacerraf problem is supposed to play.

Let me now turn to the broader relevance of this conclusion.

## 2.7 Broader Relevance

The difficulties surrounding the Benacerraf Problem are actually very general. They infect formulations of it aimed at moral realism, modal realism, logical realism, and philosophical realism. But they infect much more than this. They infect any argument which grants the (actual) truth and (defeasible) justification of our beliefs from an area and seeks to undermine those beliefs, so long as the area *F* meets two conditions. Those conditions are:

1. The *F*-truths would be metaphysically necessary.
2. There is a plausible explanation of our having the *F*-beliefs that we have which shows that we could not have easily had different ones.

Many arguments which are not supposed to be variations on the Benacerraf Problem have these features. Consider Evolutionary Debunking Arguments, which are now influential epistemological challenges themselves. Richard Joyce offers a canonical formulation:

> Nativism [the view that moral concepts are innate] offers us a genealogical explanation of moral judgments that nowhere […] presupposes that these beliefs are true […]. My contention […] is that moral nativism […] might well […] render [moral beliefs] unjustified […]. In particular, any epistemological benefit-of-the-doubt that might have been extended to moral beliefs […] will be neutralized by the availability of an empirically confirmed moral genealogy that nowhere […] presupposes their truth.
>
> Joyce (2008: 216)[44]

What is Joyce's argument? Taken at face value, it is that our moral beliefs (realistically conceived) are undermined on the mere ground that their contents fail to figure into the evolutionary explanation of our having them. But we have seen that such an argument must be too quick. In the first place, it is arguable that the contents of our moral beliefs do figure into the evolutionary explanation of our having them, just as it is arguable that the contents of our mathematical beliefs do.[45] But set this possibility aside. In order for the information to which Joyce alludes to undermine our moral beliefs, it would seem prima facie to have to give us ("direct") reason to doubt that our moral beliefs are both safe and sensitive. But, on its own, the observation that the contents of our moral beliefs fail to figure into the evolutionary explanation of our having them does not do this. If this observation has any epistemological force, it is apparently to help undercut what is arguably the only dialectically effective argument *for* the contents of our moral beliefs (realistically conceived). As Gilbert Harman writes,

> Observation plays a part in science it does not appear to play in ethics, because scientific principles can be justified ultimately by their role in explaining observations […]. [M]oral principles cannot be justified in the same way.
>
> Harman (1977: 10)

If the contents of our moral beliefs did figure into the evolutionary explanation of our having those beliefs, then those contents *could* be justified by their role in explaining observations.[46]

Perhaps, then, Joyce's argument is that the evolutionary considerations to which he alludes give us reason to doubt that our moral beliefs are safe. But, on the contrary, if anything, those considerations seem to give us reason to believe that our moral beliefs *are* safe. The whole point of Evolutionary Debunking Arguments is often taken to be that it would have been advantageous for our ancestors to have the "core" moral beliefs that we have (such as that killing our offspring is bad) "independent of their truth." If this is right, then we could not have easily come to have different such beliefs, in which case they are safe (assuming that the explanatorily basic moral truths are the same in nearby worlds—more on this below). As before, we may be able to "bootstrap up" from the safety of our "core" moral beliefs to the safety of our moral theories. Note the irony. A tentative sign that a realist about an area *F* can establish the safety of her beliefs is *that there is an Evolutionary Debunking Argument aimed at F-realism*.[47]

Perhaps, then, rather than giving us reason to believe that our moral beliefs are not safe, Joyce takes evolutionary considerations to give us reason to doubt that our moral beliefs are sensitive. This interpretation is in harmony with standard formulations of the Evolutionary Debunking Argument, due to Walter Sinnott-Armstrong, Sharon Street, Michael Ruse, and others. It would have benefited our ancestors to believe that killing our offspring is wrong even if killing our offspring were right—or, indeed, even if there were no (atomic or existentially quantified) moral truths at all. Thus, had the contents of our moral beliefs been false, we still would have believed those contents. In an earlier book, Joyce writes,

> Suppose that the actual world contains real categorical requirements — the kind that would be necessary to render moral discourse true. In such a world humans will be disposed to make moral judgments […] for natural selection will make it so. Now imagine instead that the actual world contains no such requirements at all — nothing to make moral discourse true. In such a world, humans will *still* be disposed to make these judgments… for natural selection will make it so.
>
> Joyce (2001: 163)

But this argument is also fallacious. Even setting aside the prospect of arguing for sensitivity via explanatory indispensability, the explanatorily basic moral truths —the truths that fix the conditions under which a concrete person, action, or event satisfies a moral predicate—are widely supposed to be metaphysically necessary. But, if they are, then our corresponding beliefs are vacuously sensitive on a standard semantics. Of course, as before, those beliefs may not be non-vacuously sensitive. But, then, neither are our relevantly uncontroversial beliefs, such as the belief that atoms arranged car-wise compose cars.

As in the mathematical case, it follows that we may be able to show that our moral beliefs are both safe and sensitive, given their actual truth. This affords them extraordinary intellectual security. To the extent that Joyce's evolutionary speculations give us no reason to doubt their modal security, it is hard to see how those speculations could undermine our moral beliefs.

## 2.8 The Benacerraf Problem and the Gettier Problem

What is the Benacerraf Problem? There does not seem to be a satisfying answer. There does not seem to be a sense of "explain the reliability" in which it is plausible *both* that it appears in principle impossible to explain the reliability of our mathematical beliefs and that the apparent in principle impossibility of explaining their reliability undermines them. The problem is quite general, infecting all arguments with the structure of the Benacerraf Problem meeting two conditions.

It remains open how this conclusion bears on our claim to knowledge in mathematics and related areas. Unlike Benacerraf's challenge, both Field's challenge and Evolutionary Debunking Arguments focus on whether our beliefs are justified, not on whether they qualify as knowledge. This is a virtue. The correct analysis of knowledge is notoriously controversial. Moreover, if our beliefs are justified, and we can relevantly explain their reliability, then it is hard to see why we should give them up—*even if* they fail to qualify as knowledge. Perhaps the most interesting feature of Field's formulation of the Benacerraf Problem and of Evolutionary Debunking Arguments is that they purport to give realists reason to *change their views*.

Nevertheless, the argument offered here may suggest that we have knowledge in mathematics and related areas. I have argued that our mathematical and related beliefs may be both safe and sensitive, given their (actual) truth (and defeasible justification). (In the case of mathematics, I also argued that the contents of our beliefs may figure into the best explanation of our having them.) But many philosophers would hold that a justified true belief which is both safe and sensitive qualifies as knowledge (and even more would hold that a justified true belief which is both safe and sensitive and such that its content figures into the best explanation of our having it so qualifies).

Perhaps the present discussion helps to explain why. "Gettiered" beliefs—justified and true beliefs which fail to qualify as knowledge—are plausibly beliefs whose truth is coincidental in a malignant sense. What is that sense? It is arguably precisely the sense in which learning that the truth of one's beliefs is coincidental would undermine them. If this is correct, then there is a "translation scheme" between the claim that it is impossible to relevantly explain the reliability of our $F$-beliefs, given their truth, and the claim that those beliefs are Gettiered.

## Notes

1. See Benacerraf (1973: 675).
2. *Ibid*: 673.
3. See also Mackie's epistemological "argument from queerness" in Chap. 1 of Mackie (1977), as well as Alan Gibbard's discussion of "deep vindication" in Sect. 13 of his Gibbard (2003).
4. Similarly, O'Leary-Hawthorne writes, "The relevant difficulties are not, of course, peculiar to modal metaphysics. Our dilemma re-enacts Paul Benacerraf's famous dilemma for the philosopher of mathematics" (O'Leary-Hawthorne 1996: 183).
5. See also Resnik (2000).
6. Adam Pautz suggests that the Benacerraf Problem arises for realism about phenomenal properties. He writes, "If propositions about resemblances among properties report acausal facts about abstracta, then how can we explain the following regularity: generally, if we believe *p*, and *p* is a such a proposition, then *p* is true? […] [T]his problem arises for all accounts of the Mary case [of the "Mary's Room" argument] […]. [And it] resembles the Benacerraf-Field problem about mathematics" (Pautz 2011: 392–3). (The Mary's Room argument purports to show that physicalism is false on the grounds that someone—Mary—could know all the physical facts about color vision and yet learn something upon seeing colored things for the first time.).
7. The only contemporary formulation of a causal constraint on knowledge of which I am aware seems hopelessly ad hoc. Colin Cheyne contends that "[i]f *F*s are noncomparative objects, then we cannot know that *F*s exist unless our belief in their existence is caused by: (a) an event in which *F*s participate, or (b) events in which each of the robust constituents of *F*s participate, or (c) an event which proximately causes an *F* to exist" (Cheyne 1998: 46).
8. For overviews, see Liggins (2006, 2010) and Linnebo (2006).
9. I am not actually a mathematical realist (in a common sense of that phrase). But since it will be convenient to frame things in terms of what "we" can explain, I will often identify as one.
10. I will generally fail to qualify discussion of mathematical truths or of our mathematical beliefs with "realistically conceived" in what follows. But let me emphasize that this will always be what I intend. The Benacerraf Problem threatens our mathematical beliefs *under a realist construal*. This hardly detracts from its importance, however, for it is notoriously difficult to give a satisfactory non-realist construal of the subject. I clarify the relevant sense of "mathematical realism" at the end of the present section.
11. These explanatory tasks are distinguished in Schechter (2010).
12. For more on the "non-skeptical" character of Field's challenge, see Balaguer (1995).
13. But see Sect. 3 of Clarke-Doane (2014).
14. See, for example, Shapiro (1995) or Leng (2008).

15. Joshua Schechter, following Field (1998), suggests that what matters for Field's formulation of the Benacerraf Problem is not the ontology of mathematics, but its *objectivity*. He has in mind the contrast between, say, standard platonism, according to which there is a "unique" universe of mathematical objects that our (fundamental) mathematical theories aim to describe, and Balaguer's "Full-Blooded Platonism," to be discussed below, according to which any "intuitively consistent" theory that we might have come to accept would have been true. A given domain of truths $F$ is objective in the relevant sense, if and only if "not just any [$F-$] practice counts as correct." Schechter writes, "[t]he root of the trouble is not the ontology but the apparent *objectivity* of mathematics […]. If mathematics […] were to turn out not to have an ontology, but the relevant truths were nevertheless objective, our reliability would remain puzzling" (Schechter 2010: 439). We will see in Sect. 3 that this thought, as promising as it may appear, could not be correct. (For more in-depth treatment of this issue, see Clarke-Doane [manuscript]).
16. The parenthetical qualification is needed in order to distinguish FBP, which is highly controversial, from the Completeness Theorem, which is not. FBP does not just say that every consistent mathematical theory has a model. It says that every such theory has an *intended* model. Sometimes critics of FBP accuse Balaguer of merely advocating the Completeness Theorem. See, for instance, Burgess (2001).
17. I assume the standard view that "had it been the case that $p$, it would have been the case that $q$" is true only if $q$ is true at the closest worlds in which $p$ is true.
18. While I am not using "safe" in exactly the sense of Pritchard (2005) or Williamson (2000), the idea is similar. It is unobvious how to spell out the pertinent sense of "easily." See Hawthorne (2004: 56) for a complication. I will mostly ignore methods of belief-formation in what follows.
19. I am not suggesting that our having *true "core" mathematical beliefs* per se might be evolutionary inevitable. I am suggesting that our having the "core" mathematical beliefs that we have, *which are actually true*, might be. These claims are very different. See Field (2005) and Clarke-Doane (2012).
20. Field alludes to the parenthetical antecedent in the following: "[W]e can assume, with at least some degree of clarity, a world without mathematical objects…" (Field 2005: 80–81). Sometimes Field describes the problem of showing that our mathematical beliefs are counterfactually persistent as that of offering a *unified* explanation of their reliability. In a discussion of the Benacerraf Problem for logical realism, Field writes "The idea of an explanation failing to be "unified" is less than crystal clear, but another way to express what is unsatisfactory about [a bad explanation] is that it isn't *counterfactually persistent* […], it gives no sense to the idea that if the logical facts had been different then our logical beliefs would have been different too." (Field 1996: 371). I will discuss another sense of "unified explanation" in Sect. 6.
21. This is not incontrovertible. Given an "external" theory of reference, one can argue that, had we been brains in vats, we would have believed that we were. See Putnam (1981). This argument relies on a causal theory of reference.

22. See also Pust (2004).
23. There *are* arguments that the mathematical truths would be necessary, though I am not sure how compelling they are. One such argument is that mathematical truths concern abstract objects like numbers, sets and tensors. Such objects are neither spatial nor temporal and participate in no physical interactions. "Hence" they do not depend on the way that the contingent world happens to be. So, if mathematical objects satisfy the standard axioms in the actual world, then they do so in all possible worlds. See Shapiro (2000: 21–24) for something like this argument.
24. *See* Korman (2014): Sect. 4.2 for relevant discussion.
25. This assumes a closure principle which could conceivably be questioned. But even if it were, an analogous point would hold. See Clarke-Doane (2016), Sect. 2.2.
26. Field, of course, hopes ultimately to show that mathematics is not assumed by (the best formulations of) our empirical theories. But he does not deny that it *seems* to be, and, anyway, appears to hold that the Benacerraf Problem arises even if it is. See Field (2005): Sect. 2.5, and Field (1989: 262).
27. See, for instance, Baker (2009), and Lyon and Colyvan (2008).
28. For more on evolutionary examples like this, see Braddock et al. (2012). I argue that, despite appearances, the contents of our mathematical beliefs do not explain our having them in Clarke-Doane (2012): Sect. III.
29. For more on this, see Sect. 2.3 of Clarke-Doane (2014).
30. Field makes gestures in the direction of this position in Field (1996), but then explicates "unified explanation" in terms of sensitivity (see *supra* footnote 22). Sharon Street makes claims like this with respect to our moral beliefs in Street (2006, 2008).
31. What if we only assign objective probability 1 to contents which are necessary in an even stronger sense—e.g. "conceptually necessary"? Then, unless one can argue that "ontological nihilism" is not just false but *conceptually impossible*, the contents of our (explanatorily basic) "ordinary object beliefs" would seem to have equal claim to being objectively improbable. See Sect. 2.4.
32. I am not assuming that if our mathematical beliefs are safe, then the objective probability that they are true is high. I am pointing out that, for all that has been said, our mathematical beliefs may be both safe and objectively probable.
33. Perhaps it amounts to the request to show that our mathematical beliefs are "grounded in" or "constituted by" the corresponding state of affairs (see Bengson (2015) for something like this proposal)? Such hyperintensional ideology does not seem to me to be more perspicuous than the quoted phrase itself. But, even if it were, this proposal would not seem to serve Field's purposes, as will become clear shortly.
34. Safety and sensitivity plausibly need to be relativized to methods of belief formation, as indicated towards the end of Sect. 2.3 and at the beginning of Sect. 2.4, respectively and reason is shorthand for "direct" "reason". (When the area $F$ is not logic, "all of our $F$-beliefs" refers to all of our non-logical $F$-beliefs. Not even Field would deny that we are justified in believing, e.g., that

either it is the case that there are perfect numbers greater than 1,000,000 or it is not the case that there are.).

35. The next few paragraphs closely follow parts of Sect. 2.4 of Clarke-Doane (2016).” Thanks to Neil Sinclair for permission to reprint them.
36. See Pollock (1986: 38–39) for the distinction between underminers (or “undercutters”) and rebutters.
37. Thanks to David James Barnett for pressing me with something like this example.
38. Compare to the discussion of Full-Blooded Platonism in Sect. 2.3.
39. This assumes that we believe of at least one sentence that it is invalid. But it is hard to see how we could believe of any sentence that it is valid while failing to believe of any other that it is invalid.
40. See Conee and Feldman (1998). We could alternatively individuate kinds by methods of belief. Modal Security would then say that if *E* obligates us to give up all of our beliefs formed via *M*, then it does so by giving us reason to doubt that our beliefs formed via *M* are both sensitive and safe. Given plausible assumptions, this formulation is strictly stronger than Modal Security as I have interpreted it.
41. There is another kind of problem case. Suppose that *E* is evidence that a false theory of justification is true according to which our *F*-beliefs are not justified. It might be thought that *E* could undermine all of our *F*-beliefs, but not by giving us reason to doubt their sensitivity or safety. But, on inspection, this seems bizarre. Suppose that *F* includes only propositions for which we have excellent evidence, and *E* is evidence for the view that a belief is justified only if it is infallible. Perhaps we are students in a Philosophy 101 class, for example, and *E* is an apparently strong argument for the view. To give up our *F*-beliefs on the basis of *E*—when *E* is neither “rebutting” nor “direct” reason to doubt that our *F*-beliefs are modally secure–seems to be to give up those beliefs “for the wrong kind of reason” (Barnett, [unpublished manuscript]).
42. See Field (2005) and Clarke-Doane (2012).
43. Strangely, Schechter appears to grant something like this point in his discussion of Nagel in Schechter (2010: 447–448). See also, Chap. 4 of Nagel (1997). My own view is that the interest of the Benacerraf Problem is greatly reduced if the apparent in principle impossibility of answering it is not supposed to undermine our mathematical beliefs (realistically construed).
44. In addition to Joyce, see Greene (2007), Griffiths and Wilkins (forthcoming), Levy (2006), Lillehammer (2003), Ruse (1986), Sinnot-Armstrong (2006), and Street (2006).
45. See Brink (1989), Boyd (2003a, b) for relevant discussion.
46. I argue that debunkers have confused the challenge to empirically justify our moral beliefs with the challenge to explain their reliability in Clark-Doane (2015).
47. I am not saying that Joyce and Street is committed to holding that we could not have easily had different explanatorily basic moral beliefs. I am saying that the

view that we could not have is consistent with, and on some presentations, even suggested by, their evolutionary speculations. As a result, those speculations certainly do not seem to give us reason to believe that we could have easily had different explanatorily basic moral beliefs.

## References


Baker, A. (2009). Mathematical explanation in science. *The British Journal for the Philosophy of Science, 60*(3), 611–633.

Balaguer, M. (1995). A platonist epistemology. *Synthese, 103*(3), 303–325.

Balaguer, M. (1998). *Platonism and anti-platonism in mathematics*. Oxford: Oxford UP.

Balaguer, M. (1999). Review of Michael Resnik's mathematics as a science of patterns. *Philosophia Mathematica, 7*(3), 108–126.

Bealer, G. (1999). A theory of the a priori. *Philosophical Perspectives, 13*, 29–55.

Benacerraf, P. (1973). Mathematical truth. *Journal of Philosophy, 70*, 661–679.

Bengson, J. (2015). Grasping the third realm. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 5, pp. 1–34). Oxford: Oxford UP.

Boyd, R. (2003a). Finite beings, finite goods: The semantics, metaphysics and ethics of naturalist consequentialism, part I. *Philosophy and Phenomenological Research, 66*(3), 505–553.

Boyd, R. (2003b). Finite beings, finite goods: The semantics, metaphysics and ethics of naturalist consequentialism, part II. *Philosophy and Phenomenological Research, 67*(1), 24–47.

Braddock, M., Mortensen, A., & Sinnot-Armstrong, W. (2012). Comments on Justin Clarke-Doane's. 'Morality and mathematics: The evolutionary challenge'. *Ethics Discussions at PEA Soup.* http://peasoup.typepad.com/peasoup/2012/03/ethics-discussions-at-pea-soup-justin-clarke-doanes-morality-and-mathematics-the-evolutionary-challe-1.html

Brink, D. (1989). *Moral realism and the foundations of ethics*. Cambridge: Cambridge UP.

Burgess, J. P. (2001). Review of platonism and anti-platonism in mathematics. *The Philosophical Review, 110*, 79–82.

Carnap, R. [1950] (1983). Empiricism, semantics and ontology. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics—Selected readings* (2nd ed., pp. 241–257). Cambridge: Cambridge UP.

Casullo, A. (2002). A priori knowledge. In P. Moser (Ed.), *Oxford handbook of epistemology* (pp. 95–143). Oxford: Oxford UP.

Cheyne, C. (1998). Existence claims and causality. *Australasian Journal of Philosophy, 76*(1), 34–47.

Clarke-Doane, J. (Manuscript) "Mathematical Pluralism and the Benacerraf Problem".

Clarke-Doane, J. (2016). Debunking and Dispensability. In Neil Sinclair and Uri Leibowitz (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability* (pp. 24–36). Oxford: Oxford University Press.

Clarke-Doane, J. (2012). Morality and mathematics: The evolutionary challenge. *Ethics, 122*(2), 313–340.

Clarke-Doane, J. (2014). Moral Epistemology: The Mathematics Analogy. *Noûs, 48*(2), 238–255.

Clarke-Doane, J. (2015). Justification and Explanation in Mathematics and Morality. Russ Shafer-Landau (Ed.), Oxford Studies in Metaethics, Vol. 10, pp. 80–104. New York: Oxford University Press.

Conee, E., & Feldman, R. (1998). The generality problem for reliabilism. *Philosophical Studies, 89*(1), 1–29.

Field, H. H. (1989). *Realism, mathematics, and modality*. Oxford: Basil Blackwell.

Field, H. H. (1996). The a prioricity of logic. *Proceedings of the Aristotelian Society New Series, 96*, 359–379.
Field, H. H. (1998). Mathematical objectivity and mathematical objects. In C. MacDonald & S. Laurence (Eds.), *Contemporary readings in the foundations of metaphysics* (pp. 387–403). Oxford: Basil Blackwell.
Field, H. H. (2005). Recent debates about the a priori. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 1, pp. 69–88). Oxford: Clarendon Press.
Gibbard, A. (2003). *Thinking how to live*. Cambridge, Mass: Harvard UP.
Goldman, A. I. (1967). A causal theory of knowing. *The Journal of Philosophy, 64*(12), 357–372.
Greene, J. (2007). The secret joke of Kant's soul. In W. Sinnot-Armstrong (Ed.), *Moral Psychology, Vol. 3: The neuroscience of morality: Emotion, disease, and development* (pp. 35–80).
Griffiths, P. & Wilkins, J. Forthcoming. When do evolutionary explanations of beliefs debunk belief? PhilSci Archive: http://philsci-archive.pitt.edu/5314/
Hamkins, J. (2012). The set-theoretic multiverse. *Review of Symbolic Logic, 5*, 416–449.
Harman, G. (1977). *The nature of morality: An introduction to ethics*. Oxford: Oxford UP.
Hart, W. D. (1977). Review of *Mathematical Knowledge*. *The Journal of Philosophy, 74*(2), 118–129.
Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford UP.
Huemer, M. (2005). *Ethical intuitionism*. New York: Palgrave Macmillan.
Joyce, R. (2001). *The myth of morality*. Cambridge: Cambridge UP.
Joyce, R. (2008). Précis of the evolution of morality [and reply to critics]. *Philosophy and Phenomenological Research, 77*(1), 213–218 [pp. 218–267 for the "Reply to Critics"].
Korman, D. Z. (2014). Ordinary objects. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2014/entries/ordinary-objects/
Leng, M., Paseau, A., & Potter, M. (Eds.). (2008). *Mathematical knowledge*. Oxford: Oxford UP.
Levy, N. (2006). Cognitive scientific challenges to morality. *Philosophical Psychology, 19*(5), 567–587.
Lewis, D. (1986). *On the plurality of worlds*. Oxford: Wiley-Blackwell.
Liggins, D. (2006). Is there a good epistemological argument against platonism? *Analysis, 66* (290), 135–141.
Liggins, D. (2010). Epistemological objections to platonism. *Philosophy Compass, 5*(1), 67–77.
Lillehammer, H. (2003). Debunking morality: Evolutionary naturalism and moral error theory. *Biology and Philosophy, 18*(4), 567–581.
Linnebo, Ø. (2006). Epistemological challenges to mathematical platonism. *Philosophical Studies, 129*(3), 545–574.
Linsky, B., & Zalta, E. (1995). Naturalized platonism versus platonized naturalism. *The Journal of Philosophy, 92*(10), 525–555.
Lyon, A., & Colyvan, M. (2008). The explanatory power of phase spaces. *Philosophia Mathematica, 16*(2), 227–243.
Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. New York: Penguin.
Nagel, T. (1997). *The last word*. Oxford: Oxford UP.
Nozick, R. (1981). *Philosophical explanations*. Cambridge, Mass: Harvard UP.
O'Leary-Hawthorne, J. (1996). The epistemology of possible worlds: A guided tour. *Philosophical Studies, 84*(2–3), 183–202.
Pautz, A. (2011). Can disjunctivists explain our access to the sensible world? *Noûs* (*Supplement: Philosophical Issues, Epistemology of Perception*), *21*, 384–433.
Peacocke, C. (1999). *Being known*. Oxford: Oxford UP.
Pollock, J. (1986). *Contemporary theories of knowledge*. Lanham, Maryland: Rowman and Littlefield.
Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford UP.
Pust, J. (2004). On explaining knowledge of necessity. *Dialectica, 58*(1), 71–87.
Putnam, H. (1981). *Reason, truth and history*. Cambridge: Cambridge UP.

Putnam, H. [1980] (1983). Models and reality. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics—Selected readings* (2nd ed., pp. 421–444). Cambridge: Cambridge UP.
Resnik, M. (2000). Against logical realism. *History and Philosophy of Logic, 20*(3–4), 181–194.
Ruse, M. (1986). *Taking Darwin seriously*. Amherst: Prometheus Books.
Schechter, J. (2010). The reliability challenge and the epistemology of logic. *Noûs* (*Supplement: Philosophical Perspectives*), *24*, 437–464.
Schechter, J. (2013). Could evolution explain our reliability about logic? In J. Hawthorne & T. Szabò (Eds.), *Oxford studies in epistemology* (Vol. 4, pp. 214–239). Oxford: Oxford UP.
Shapiro, S. (1995). Modality and ontology. *Mind, 102*(407), 455–481.
Shapiro, S. (2000). *Thinking about mathematics: Philosophy of mathematics*. Oxford: Oxford UP.
Sinnott-Armstrong, W. (2006). *Moral skepticisms*. Oxford: Oxford UP.
Stalnaker, R. (1996). On what possible worlds could not be. In *Ways a world might be: Metaphysical and anti-metaphysical essays* (pp. 40–54). Oxford: Oxford UP.
Steiner, M. (1973). Platonism and the causal theory of knowledge. *The Journal of Philosophy, 70* (3), 57–66.
Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies, 127*(1), 109–166.
Street, S. (2008). Reply to Copp: Naturalism, normativity, and the varieties of realism worth worrying about. *Philosophical Issues (Interdisciplinary Core Philosophy), 18,* 109–166.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford UP.

# Chapter 3
# Benacerraf's Mathematical Antinomy

**Brice Halimi**

## 3.1 Benacerraf's Dilemma About Mathematics as a Mathematical Antinomy

Benacerraf's "Mathematical Truth" (Benacerraf 1973) takes on the form of a well-known dilemma. Either a referential semantics for ordinary language is extended to mathematical language, but then one lapses into platonism, or a reasonable account of mathematical knowledge as a proof activity is put forward, but then no account of mathematical truth as truth is given. Of course, the semantic horn and the epistemic horn are not exactly on a par, but Benacerraf's paper precisely tends to tilt the scales so as to get an embarrassing balance, i.e. a dilemma. Its two horns are mutually exclusive and, to the extent that they are both indefensible, Benacerraf leaves it to the reader to understand that they both rest on a mistake.

I shall argue that the situation smacks of the paradigmatic opposition of two equally false theses that constitutes the core of Kant's antinomies of pure reason—more precisely the core of the first two antinomies, the so-called "mathematical antinomies," whose both opponents, Kant argues, are mistaken, whereas both opponents of the last two antinomies (the "dynamical" ones) are right, albeit from two incompatible points of view. It might seem at first sight that Benacerraf's dilemma is a dynamical antinomy, as though mathematical truth could be looked at from the point of view of epistemology as well as from that of semantics, so that both claims would be legitimate within their respective limits. It is nevertheless clear that Benacerraf's dilemma must be understood as an antinomy of the mathematical kind, where both opposite claims are false, since Benacerraf makes it plain that neither is philosophically defensible.

B. Halimi (✉)
Université Paris Nanterre (IREPH) & SPHERE (France), Paris, France
e-mail: bhalimi@u-paris10.fr



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_3

A comparison of Benacerraf's dilemma with Kant's mathematical antinomies is called for by strong analogies, but it also serves a purpose. It is indeed natural to ask what Benacerraf's dilemma is driving at, since Benacerraf himself does not provide any clear solution to it. On that score, a comparison with Kant's Transcendental Dialectic could turn out to be useful since Kant, in addition to presenting a predicament analogous to Benacerraf's dilemma, does provide a solution for it. The aim of the present paper is to go into the details of the analogies so as to explore the possibility of transposing Kant's solution to the case of Benacerraf's dilemma.

Let us recall the first two antinomies of pure reason. The first lies in the conflict between two theses. The first thesis of the first antinomy is "The world has a beginning in time, and in space it is also enclosed in boundaries." Its antithesis is "The world has no beginning, and no bounds in space, but is infinite with regard to both time and space."[1]

As to the second antinomy, its thesis is: "Every composite substance in the world consists of simple parts, and nothing exists anywhere except the simple or what is composed of simples." Its antithesis claims: "No composite thing in the world consists of simple parts, and nowhere in it does there exist anything simple."[2]

In Kantian antinomies, each claim is mainly negative: it relies on a *reductio ad absurdum* and feeds entirely upon the impossibility of the opposite claim. In the same way, each horn of Benacerraf's dilemma draws its strength from the predicament of the other only. So the main structure of the argumentation is the same in both cases.

Let us now examine the particulars of the antinomies of pure reason. There is nothing accidental about the antinomies—nor indeed about the whole dialectic of pure reason. Pure reason is bound to run into contradictions as soon as it is applied to objects of experience. Indeed, a concept of reason (such as the concept of the world) seeks what is unconditioned with respect to some given condition and, for that purpose, carries out the synthesis of the regressive series of conditions for a given conditioned or, rather, takes that synthesis for granted (i.e. already completed) and presupposes the absolute totality of the series of the conditions. So there are in fact two different ways of conceiving of the unconditioned[3]: either as being the last term of a regressive series of conditions, or as consisting in the whole series itself.

In each antinomy, the thesis encapsulates the first conception, the antithesis the second one. The thesis seeks a first unconditioned entity on which the whole series of conditions depends: it is reason trying to catch up with understanding, since the unconditioned is presented as an actual object (although it is prevented from being an empirical one). Conversely, the antithesis presents the absolute sum of the conditions in the series itself as constituting an unconditioned totality: it is understanding trying to catch up with reason, since it extends the successive synthesis of appearances (empirically discharged by the understanding) into an ideally completed synthesis of the whole series of conditions. As a consequence, cosmological ideas are either too large or too small for the empirical regress sustained by the concepts of the understanding. The first antinomy offers an example of this discrepancy:

> [Assume] that *the world has no beginning*: then it is too big for your concept; for this concept, which consists in a successive regress, can never reach the whole eternity that has elapsed. Suppose *it has a beginning*, then once again it is *too small* for your concept of understanding in the necessary empirical regress. For since the beginning always presupposes a preceding time, it is still not unconditioned, and the law of the empirical use of the understanding obliges you to ask for a still higher temporal condition, and the world is obviously too small for this law.
>
> Kant [1787] (1998: A486-487/B514-515; 508–509)

Hence each antinomy originates in the conflict between understanding and reason. In each case, the antithesis sticks to the limits of sensible experience given by the conditions of space and time. On the contrary, the thesis goes beyond those limits and aims at some given absolute entity whose purported existence stems from a request of reason to make sense of a totality of conditions, starting with some given conditioned. Owing to these orientations, Kant associates the thesis of each antinomy with dogmatism and the corresponding antithesis with empiricism.[4] With empiricism embodied in the antithesis, "the understanding is at every time on its proper ground, namely the field solely of possible experiences" (Kant op. cit.: A468/B496; 499): the connection between appearances and the laws of those connections is the main focus. On the contrary, the dogmatism embodied in the thesis expresses the voice of reason in its quest for an unconditioned entity *in individuo*—what Kant calls an *Objekt*, as opposed to a *Gegenstand*.

At this point, it seems fair to say that an analogy with Benacerraf's dilemma is called for. How far that analogy goes has to be examined, but it is natural to suspect an analogy between the stress laid by empiricism on the understanding and the stress laid by the epistemic horn on knowledge; and between the stress laid by dogmatism on *Objekte* and the stress laid by platonism on mathematical objects. Of course, both contexts differ substantially. But, after all an analogy is not a comparison, quite the contrary. The whole working hypothesis of this paper, far from proposing an argument in favor of any direct resemblance, is that there is an analogy relating, on the one hand, the function of the understanding to the function of mathematical proofs and, on the other, the function of entities of pure reason to the function of mathematical objects. As one shall see, the analogy goes deep and leads to a solution to Benacerraf's dilemma.

Kant's *Critique* characterizes the understanding as the faculty of relations and laws, and reason as the faculty of absolute entities. In the same way, the epistemic horn of Benacerraf's dilemma belongs in the camp of formal proofs and syntactic rules, whereas the semantic horn is typical of a quest for an absolute mathematical referent. Empiricism, Kant says, encourages and furthers knowledge, but at the cost of the practical (in the Kantian sense), whereas dogmatism, especially platonism, meets the practical interest of reason,[5] but neglects the investigation of natural appearances. In the same way, the epistemic horn focuses on proofs and on mathematical activity taken in itself, at the cost of mathematical referentiality, whereas the semantic horn meets the practical need of a unified referential framework, but lets the objects prevail over one's possible access to them.

Thus, the analogy leading from the antinomy of pure reason to Benacerraf's dilemma consists in a comparison of Kantian understanding with the production of

deductive series, and of Kantian reason with the reference to the absolute entities underlying those series; in a comparison of the platonism of the thesis with Gödelian platonism, and of the empiricism of the antithesis with Hilbertian finitism. Instead of claiming that space has a definite extension, the dogmatic in Benacerraf claims that mathematical objects are the definite actual denotations of mathematical terms. To go back to the first antinomy, instead of claiming that space is boundless, the empiricist claims that a mathematical object is nothing but the unlimited sum of all the formal proofs that we can produce with respect to the symbols that stand for it, and that it does not exist over and above these proofs.[6]

There are, as a matter of fact, various and quite precise similarities between the respective situations described by Kant and Benacerraf which urge us to reconsider "Mathematical Truth" as putting forward a genuine Dialectic of Mathematical Reason. The table below summarizes these main similarities.

Analogy between the antinomy of pure reason and Benacerraf (1973)

| Kant | Benacerraf |
|---|---|
| Mathematical antinomy | Dilemma |
| Antithetic (each thesis feeds upon the contradiction of the other) | Negative argumentation |
| Understanding | "Our ability to produce and survey formal proofs" (Benacerraf 1973: 668) |
| Appearances | Symbols |
| Possible experience | Admissible inference |
| Series of conditions | Deductive chains |
| Law | Theorem |
| Reason | Truth theory |
| Unconditioned being | Direct reference |
| Intellectual intuition | Gödelian intuition |
| Dogmatism (thesis) | Realism (standard conception) |
| Platonism | Platonism (in mathematics) |
| Empiricism (antithesis) | Finitism (combinatorial view) |
| Epicureanism | Hilbertian formalism |
| The Idea is "either too big or too small" for the concept of the understanding (Kant [1787] 1998: A486/B514; 508) | All analyses "bulge either on the side of knowledge or on the side of truth" (Benacerraf 1973: 668). |
| The antithesis favors knowledge of nature (Kant [1787] 1998: A469-472/B497-500; 499–501) | The epistemic horn accounts for mathematical knowledge. |
| The thesis favors the practical and is more popular for that reason (Kant [1787] 1998: A466-467/B494-495; 498–499) | The semantic horn accounts for mathematical truth and dovetails with the pragmatic needs of ordinary communication.[7] |
| Solution provided by transcendental idealism | Solution to be specified |

For lack of space, a further study of Kant's Transcendantal Dialectic has to be foregone for the moment, but the detour through Kant calls forth one important point: Kant provided a systematic and extensive solution for his antinomies. In that sense, the detour hints at a solution of Benacerraf's dilemma. As a matter of fact, Kant's solution is in some respects reconfigured by Benacerraf himself, although not in Benacerraf (1973). I shall argue that the reconfiguration may be recovered from two other papers by Benacerraf: Benacerraf (1965) ("What Numbers Could Not Be") and Benacerraf (1981) ("Frege: The Last Logicist"). Before turning towards the possible transposition of Kant's solution to Benacerraf's problem, I shall therefore examine these papers. They will confirm that Benacerraf's dilemma is indeed akin to a mathematical antinomy, as opposed to a dynamical one, i.e. that the two views in conflict cannot be reconciliated and must be both overcome. I shall then consider the transposition of Kant's solution, and to what extent Benacerraf carries it out himself.

## 3.2 Bring on the Kids and the Founding Father

This section aims to show that Benacerraf (1965) proves the semantic horn to be inconsistent and, proceeding forward, that Benacerraf (1981) proves the epistemic horn to be inconsistent.

### *3.2.1 Backwards to Benacerraf 1965: Objects Involve Proofs*

If Benacerraf (1973) foregrounds the semantic constraint, Benacerraf (1965) clearly shows that the semantic values of numerical terms like "3" or "17" are not univocal. Even when we are using genuine singular terms in mathematics, their reference is not set unambiguously. So the basic lesson to be learned is that neither Ernie's account nor Johnny's account can provide the right semantic value for natural numbers. On that score, neither Ernie nor Johnny succeeds in securing any of the two horns.

Actually, Ernie's account (i.e. Zermelo's) and Johnny's (i.e. von Neumann's) can be construed as two different interpretations of the same mathematical objects. A mathematical object such as 3 is in fact a series of objects (containing $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$ as well as $\{\{\{\varnothing\}\}\}$), supplemented with the proof that the two models to which they respectively belong are *mutually interpretable*.

Let's consider $E$ and $J$, respectively the "Ernie interpretation" and the "Johnny interpretation" of $L = \{\in, \underline{0}, S\}$: $E$ is the $L$-structure whose domain is $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \ldots\}$ and $J$ is the $L$-structure whose domain is $\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \ldots\}$.

An *interpretation* of $J$ in $E$ consists:

- in a formula $\partial(x)$ of $L$
- for each atomic formula $\varphi(y_0,\ldots, y_{m-1})$ of $L$, in a formula $\varphi^*(x_0,\ldots, x_{m-1})$ of $L$
- in a surjective map $f: \partial^{\mathrm{E}} \subseteq |E| \rightarrow |J|$ such that, for all atomic formulae $\varphi(x_0,\ldots, x_{\mathrm{m}-1})$ of $L$ and all $a_{\mathrm{i}} \in \partial^{\mathrm{E}}$ $(0 \leq i \leq m - 1)$:

$$E| = \varphi^*[a_0, \ldots, a_{m-1}] \text{ iff } J| = \varphi[f(a_0), \ldots, f(a_{m-1})].$$

The formula $\partial$ is called the *domain formula* of the interpretation, and the map $f$ is called its *coordinate map*: it assigns to each element $f(a)$ in $|J|$ its "coordinate" $a$ in $|E|$.

There is a straightforward interpretation of $J$ in $E$, given by:

- $\partial(x)$ is "$x = x$," so that $\partial^E = |E|$
- $f(\underline{0}^E) = \underline{0}^J$ and $f(n \cup \{n\}) = \{n\}$, so that $f((S(\underline{n}))^E) = (S(\underline{n}))^J$
- $(x \in y)^* = x \in^* y$, i.e. $x \in \mathrm{TC}(y)$, where $\in^*$ is the ancestral relation of $\in$ and "TC($y$)" refers to the transitive closure of $y$.

(Let ZFC$^+$ be the theory ZFC extended by the definition of the function symbol "TC($y$)"; ZFC$^+$ is a conservative extension of ZFC—any model of ZFC can be expanded to a model of ZFC$^+$.)

More generally, $\varphi^* = \varphi[\in^* / \in]$. One can check that $E \models x \in y\ [n, p]$ iff $J \models x \in^* y\ [n, p]$. For instance, $E \models \underline{0} \in \underline{2}$, and $J \models \underline{0} \in^* \underline{2}$, since $J \models (\underline{0} \in \underline{1} \wedge \underline{1} \in \underline{2})$.

So, as a matter of fact, natural numbers correspond to a series of models, *supplemented with the proof that any two models of the series are mutually interpretable*. This is the way to exhibit "the natural numbers" as the invariant of the series. If we have canonical interpretations (canonical coordinates) between any two models, then we can speak in the same way of the series of all presentations of "the number 3." Quite generally, a mathematical object is but the invariant of the series of all its possible realizations or "presentations,"[8] which involves in a crucial way the *proof* of their mutual interpretability into one another. The proof is built into the mathematical object *qua* mathematical. So the semantic horn is inconsistent.

Further misgivings about the semantic horn should be mentioned. First of all, explaining the referentiality of a mathematical theory, i.e. granting that the theory is not simply wheelspinning, does not require one to abide by the superficial grammar. And Benacerraf knows this very well.[9] But that issue can be left aside, to the extent that the problem of the semantic value remains in the end an open one. A more serious concern is that the semantic constraint falls to the philosopher of mathematics. It is incumbent upon him, as a philosopher, to explain the semantic value of mathematical terms in order to account for the truth of mathematical sentences modelled upon some kind of correspondence (in the framework of Tarski's convention T). Now, the problem is a very simple one: Tarskian semantics, which is explicitly taken as the scheme to comply with, relies heavily on a background set-theoretic universe that is taken for granted. The inhabitants of that universe are

mathematical objects. In other words, Tarskian semantics is mathematical in nature. Admittedly, in the sentence "Socrates is Plato's master," "Socrates" is interpreted as denoting the individual Socrates himself, not any mathematical entity whatsoever. But what about "All men are mortal"? In Tarskian semantics, the interpretation of such a sentence calls at least for the domain of all men, which is a set—with all human beings as urelemente. So it is not true that, in the perspective of the semantic horn, mathematical objects come into play for the interpretation of mathematical terms *only*: they come into play for the interpretation of *any* sentence whatsoever. Mathematical objects are resorted to, not only to account for mathematical truth, but, more generally, to account for truth *simpliciter*.

Now the actual semantic values of the set-theoretic terms that any Tarskian-style semantics resorts to *mus*t be accounted for themselves. The mathematical terms of the semantic metalanguage ought to be properly interpreted. So the problem is only pushed back one step further; one is clearly up for an infinite regress. And since Benacerraf's epistemological point is about *any* kind of mathematical object, this problem applies not only to set theory strictly construed, but to all branches of mathematics—to arithmetic, to begin with. Worse, that infinite regress is of the vicious kind. One may indeed be reminded of Russell's remark: "[…] whenever the *meaning* of a proposition is in question, an infinite regress is objectionable, since we never reach a proposition which has a definite meaning."[10]

### 3.2.2 *Forward to Benacerraf 1981: Proofs Involve Objects*

Let us now turn to Benacerraf's later argument to the effect that the epistemic horn is inconsistent. In *Grundlagen* (§3), Frege claims that "the question [as to whether a proposition is a priori or not] is removed from the sphere of psychology, and assigned, if the truth concerned is a mathematical one, to the sphere of mathematics." Here is Benacerraf's comment:

> Since arithmetical propositions are at issue, the question of their justification is properly a matter for mathematics. Therefore, the concepts will be so defined as to make it a properly *mathematical* question whether some arithmetical judgment is analytic or synthetic, a priori or a posteriori. […] To determine whether a proposition is analytic, look for a proof of it in which the basic propositions are "primitive truths" — propositions which themselves have no proofs. If there exists such a proof (one in which appeal is made only to definitions and to "primitive truths") and the primitive truths evoked include only laws of logic, the proposition in question is analytic. If not, it is synthetic.
>
> Benacerraf (1981: 26–27)

Benacerraf extends Frege's claim with the conclusion that the whole enterprise of the *Grundlagen* is "first and foremost a mathematical one" (*Benacerraf op.cit.:* 34). Along with that assessment of Frege's perspective, it is also important to mention, with respect to §3 of the *Grundlagen*, that the question as to whether a given mathematical proposition is analytic becomes a mathematical one, not only because

"the truth concerned is a mathematical one," but because the way of establishing its analyticity is mathematical. The construction and survey of all the steps of a mathematical proof is a mathematical task—ideography being quite the tool geared to it. So only a mathematician *qua* mathematician is able to deem a given proposition analytic or synthetic. The rigorous representation of a mathematical proof is in itself a piece of mathematics—this is the implicit part of Frege's claim that "the question is assigned to the sphere of mathematics."

Benacerraf insists that Frege's theory is a theory of analytic truth. The further point that should be made here is that Frege's theory is precisely a mathematical theory laid out in the language of ideography. The recognition of special mathematical truths as analytic requires one to turn proofs themselves into mathematical objects in their own right. This is also the reason why the hierarchy of truths is, in Frege's reckoning, entirely objective. So the conclusion, if one remains true to Frege's perspective, is that mathematical proofs, even when viewed as purely formal, must be recognized as mathematical objects in their own right, i.e. as objects which are no less epistemologically problematic than the number 3 itself. To that extent, the epistemic horn is indeed inconsistent.

In his discussion of the status of Frege's definitions, Benacerraf adds a further point. With respect to the fact that, in Frege's *Grundgesetze*, any course of values can be taken to be the True, Benacerraf writes:

> Of course it does not make any mathematical difference. But *that* it makes no mathematical difference is an important philosophical point concerning what we must construe definitions such as Frege's to accomplish. Although I cannot pursue the matter further here, I hope that these examples make it clear that a straightforwardly "realist" construal of Frege's intentions or accomplishments will fail to do justice to his practice.
>
> Benacerraf op. cit.: 31

If one connects this argument with the discussion of the accounts of numbers given by Ernie and Johnny, two kinds of situations emerge where one has to deal with two mathematically equivalent items: either, as in the case of Ernie and Johnny, with two different tokens of the same mathematical character (to use Kaplan's terminology) or, as in the case of the referent of the True, with two different choices that amount to the same thing mathematically speaking. In both cases, a mathematical referent emerges in the form of a cluster of proofs proving that one is dealing with distinct but mathematically equivalent presentations of what must then be recognized as a mathematical invariant.

In his 1965 paper, Benacerraf does not intend to build up a predicament. On the contrary, and in a quite positive way, he shows that a mathematical object cannot but be the invariant of its different presentations; as such it involves the proof that any two such presentations are equivalent (in the relevant sense of "equivalent"), the equivalence proof being built into the object properly understood as a mathematical object. In his 1981 paper, Benacerraf stresses that a mathematical definition need preserve neither the meaning nor even the reference of the *definiendum*, because it is meant to introduce a mathematical object whose identity conditions are

determined by what one needs to prove, either with it or about it. In both cases, the dichotomy between mathematical objects and formal proofs, that is instrumental in setting up Benacerraf's dilemma, has to be overcome. Above all, the fact that a mathematical object consists in the system of all its different *presentations* (I shall soon go back to this word) and can never be referred to independently of one of these, is fully recognized by Benacerraf, and acknowledged by him in a much more positive way than is usually made out.

## 3.3 Back to Kant

### *3.3.1 The Analogy Put to Work*

One of the main benefits that can be expected from an analogy between Kant's Dialectic and Benacerraf's dilemma is a clue for a solution to the latter. What is Kant's solution to the whole antinomy of pure reason? Let us consider again the case of the first antinomy:

> If one regards the two propositions "The world is infinite in magnitude," "The world is finite in magnitude," as contradictory opposites, then one assumes that the world (the whole series of appearances) is a thing in itself. […] But if I take away this presupposition, or rather this transcendental illusion, and deny that it is a thing in itself, then the contradictory conflict of the two assertions is transformed into a merely dialectical conflict, and because the world does not exist at all (independently of the regressive series of my representations), it exists neither as *an in itself infinite* whole nor as *an in itself finite* whole. It is only in the empirical regress of the series of appearances, and by itself it is not to be met with at all. […] Accordingly, the antinomy of pure reason in its cosmological ideals is removed by showing that it is merely dialectical and a conflict due to an illusion arising from the fact that one has applied the idea of absolute totality, which is valid only as a condition of things in themselves, to appearances that exist only in representation, and that, if they constitute a series, exist in the successive regress but otherwise do not exist at all.
>
> Kant [1787] (1998: A504-505/B532-533; 518–519)

As a consequence, the series of conditions for a given conditioned should be understood, not as a regress *in infinitum* (a regress to infinity), but as a regress *in indefinitum* (an indeterminately continued regress) whose absolute completion cannot be postulated:

> [The regress in the series of conditions] is a principle of the greatest possible continuation and extension of experience, in accordance with which no empirical boundary would hold as an absolute boundary; thus it is a principle of reason which, as a *rule*, postulates what should be effected by us in the regress, but *does not anticipate* what is given in itself *in the object* prior to any regress.
>
> Kant [1787] (1998: A509/B537; 520)[11]

The whole antinomy, as well as the entire Dialectic, relies on the false assumption that the objects to which the ideas of reason refer are given *in themselves*, whereas they are given only in the course of a regressive series of conditions:

> [If] it is said that the world is either infinite or finite (not infinite), then both propositions could be false. For then I regard the world as determined in itself regarding its magnitude, since in the opposition I not only rule out its infinitude, and with it, the whole separate existence of the world, but I also add a determination of the world, as a thing active in itself which might likewise be false, if namely, the world were *not* given *at all as a thing in itself*, and hence, as regards its magnitude, neither as infinite nor as finite.
>
> Kant [1787] (1998: A504/B532; 517–518)

Now, if we are willing to pursue the analogy suggested at the beginning of this paper, the dilemma urges us to jettison the presupposition that mathematical objects are things in themselves and that the question of their status calls for a clear-cut answer, namely that they either "exist in the sense of a full-fledged existence" or "are mere ideal fictions in the course of a proof and do not exist at all." Let us try to implement this idea in a more precise way. How are we to make sense of the claim that mathematical objects are not given in themselves? What does it mean? It means something quite simple, namely that a mathematical object goes hand in hand with *modes of presentation* (or *presentations*, for short) whose nature depends on the kind of object it is.

Various examples of multiple modes of presentation abound in mathematics: a vector space is usually presented through an affine space fixed by the arbitrary choice of some origin; the symmetric group $S_n$ is often described as the group of permutations on $\{1, 2, \ldots, n\}$, but can equally well be defined as the group of permutations on any other $n$-element set; complex numbers can be defined algebraically as $\mathrm{R}/(X^2 + 1)$, arithmetically as the set $\{a + ib: a, b \in \mathrm{R}\}$ endowed with addition and multiplication, and geometrically as points of the plane; the natural numbers can be defined either *à la* Ernie or *à la* Johnny. Another example is Cantor's account of ordinals as equivalence classes of well-orderings: it constitutes a certain presentation of the concept of ordinal which itself involves, for each ordinal, a presentation in the form of a representative of this ordinal as equivalence class.

Admittedly, different scales ought to be distinguished, because in some cases the "same" structure is introduced with the help of different supports (as in the case of complex numbers), whereas in others different structures provide different accounts of the "same" mathematical concept (as in the case of Ernie vs. Johnny). All those cases belong to such obviously different scales that the notion of presentation is affected by some irreducible vagueness, which verges on equivocity. For instance, the different presentations of $S_n$ which have been mentioned are notational variants, whereas the different presentations of complex numbers are based on altogether different frameworks. This does not detract, however, from the occurrence of the same phenomenon on each scale, namely the diffraction of a "same" mathematical item into various incarnations, without which this mathematical item cannot be grasped, let alone studied. Such is already, in fact, Benacerraf's diagnosis in Benacerraf (1965):

> Any purpose we may have in giving an account of the notion of number and of the individual numbers, other than the question-begging one of proving of the right set of sets that *it* is the set of numbers, will be equally well (or badly) served by any one of the infinitely many accounts satisfying the conditions we set out so tediously.
>
> Benacerraf (1965: 62)

A mathematical object (e.g., a vector space) cannot in general be given "in itself," canonically, but requires some fixed configuration (e.g., an affine space having that vector space as direction) whose bias enables one to each the intended structure. Bringing up the notion of mathematical presentation is a general way to single out the pervasive use of such configurations throughout mathematics. Contrary to the actual drawing compared to the geometrical theorem, presentations do not boil down to sub-mathematical conditions of concrete mathematical activity. They are truly mathematical in nature and sometimes lend themselves to an explicit mathematical treatment: the notion of presentation of a group by generators and relations, or the notion of resolution of a module, are prominent examples in the field of algebra.

Hence, if we try to make sense of Kant's solution and to transpose it to Benacerraf's dilemma, we are lead to the view that a mathematical object is but the invariant of the *open-ended* series of all its possible presentations, which themselves are neither purely formal items nor independent semantic units.

As already pointed out at the end of Sect. 3.2.2, Benacerraf knew this very well as early as 1965:

> Number theory is the elaboration of the properties of *all* structures of the order type of numbers. The number words do not have single referents. […] Only when we are considering a particular sequence as being, not the numbers, but *of the structure of the numbers* does the question of which element is, or rather *corresponds to*, 3 begin to make any sense. Slogans like "Arithmetic is about numbers," "Number words refer to numbers," when properly urged, may be interpreted as pointing out two distinct things: (1) that number words are not names of special non-numerical entities, like sets, tomatoes, or Gila monsters; and (2) that a purely formalistic view that fails to assign any meaning whatsoever to the statements of number theory is also wrong.
>
> Benacerraf op. cit.: 70–71

Just as the cosmological idea, in Kant, "is only in the empirical regress of the series of appearances, and by itself […] is not to be met with at all" (Kant [1787] 1998: A505/B533; 518),[12] in the same way, a mathematical structure never exists outside the series of its presentations, none of which can be privileged as giving what would seem to be "the structure itself." And just as the unreachable completion of a series of conditions is the task prescribed to the understanding as a "regulative principle of reason," in the same way the never-ending exploration of all possible presentations of the "same" mathematical structure (definitions becoming theorems, and conversely), which amounts to the establishment of all the possible theorems about it, is the *task*, never amenable to completion, that defines mathematics as a discipline.

Just as "the world" is something real indeed, but only as an incomplete series of conditions and as the regulative focus thereof on "the world," in the same way a mathematical object covers an open-ended bundle of provably equivalent presentations (i.e. an open-ended bundle of equivalence proofs for different presentations), the mathematical object "itself" being only the regulative invariant of this series. It is not a thing-in-itself from which each presentation would originate, but, on the contrary, the focal point that multiple presentations project collectively as their common target.

*The presentational account* defended in this paper claims that mathematical objectivity is intentional, insofar as a coherent bundle of presentations points to an object, but only in a regulative way, without positing any pointee. On the contrary, the semantical account claims that a pointing presupposes a pointee that exists in itself, whereas the combinatorial account denies in the first place that there is any pointing at all and claims that a series of presentations, in and of itself, is all there is. Both accounts miss the crucial fact that doing mathematics involves shifting from one presentation to another (probably equivalent) one. Not only are the history and practice of mathematics witnesses to this fact, but the fact constitutes the very core and texture of mathematics.

### *3.3.2 Presentations*

A few words are in order about the notion of presentation as it occurs in the phrase "mode of presentation." The phrase may be traced back to Frege. Indeed, as is well known, it was mentioned by Frege in the context of the distinction he drew between sense and reference:

> If the sign "*a*" is distinguished from the sign "*b*" only as object (here, by means of its shape), not as sign (i.e. not by the manner in which it designates something), the cognitive value of $a = a$ becomes essentially equal to that of $a = b$, provided $a = b$ is true. A difference can arise only if the difference between the signs corresponds to a difference in the mode of presentation of that which is designated. Let *a*, *b*, *c* be the lines connecting the vertices of a triangle with the midpoints of the opposite sides. The point of intersection of *a* and *b* is then the same as the point of intersection of *b* and *c*. So we have different designations for the same point, and these names ("point of intersection of *a* and *b*", "point of intersection of *b* and *c*") likewise indicate the mode of presentation [*Art des Gegebenseins*]; and hence the statement contains actual knowledge. It is natural, now, to think of there being connected with a sign (name, combination of words, letter), besides that to which the sign refers, which may be called the reference of the sign, also what I should like to call the *sense* of the sign, wherein the mode of presentation is contained.
>
> Frege [1892] (1952: 57)

The choice that Frege made of the term "presentation" is certainly not insignificant and could be driven back to two opposite sources contemporary with Frege's work: the work of Franz Brentano in psychology, in particular his *Psychologie vom empirischen Standpunkt* (Brentano 1874), where the notion of

"mode of presentation" (*Modus des Verstellens*) is ubiquitous, most notably in his account of time-consciousness; and the foundation of the theory of group presentations by Walther von Dyck (a student of Felix Klein) in von Dyck (1882). Admittedly, the terms are different: Brentano uses "*Vorstellen*" (or "*Vorstellung*") and von Dyck uses "*Präsentation*," whereas Frege uses "*Gegebensein*." Notwithstanding these terminological differences, the affinity between the concepts put forward is strong enough to allow a comparison. In that perspective, the Fregean notion of mode of presentation should be construed as both epistemological and logical, i.e. both as an epistemic access and as a presentation in the mathematical sense of the presentation of a group. The mathematical example chosen by Frege in the quote must be understood as a way to introduce a notion that provides a common platform to account both for the basic phenomenon of meaning in ordinary language and the customary use of various descriptions of mathematical objects. But this is matter for another paper.[13]

The main point here is that, as early as 1965, before the dilemma was properly coined, Benacerraf already had a virtual solution to his 1973 dilemma, even though he did not bring it out and that his solution is as a matter of fact analogous to that proposed by Kant, according to the general analogy described at the beginning of this paper. Going back to "Mathematical Truth," one may express the mistake of the epistemic horn as the false claim that mathematical presentations do not present anything, and the mistake of the semantic horn—the mistake of contemporary structuralism[14]—as the false claim that presentations are mere artefacts, as opposed to the mathematical structures "themselves."

Of course, a lot remains to be explained about the workings of mathematical presentations and about their epistemic accessibility. I cannot provide such an explanation within the limits of this paper. For sure, mathematical presentations take on various aspects, from the choice of letters for the representation of permutations or Gödel numberings, to group presentations or module resolutions. Their spectrum is hardly amenable to a single kind and, more importantly, as underlined earlier in Sect. 3.3.1, their different cases belong to heterogeneous scales. Moreover, presentations are certainly many-layered: the presentation of some mathematical object can itself be turned into an object with respect to some of its own presentations.

Despite this variety, a mathematical presentation always relies on a fixed configuration, laid out in an environment that shows how shifting to another configuration would be possible but incidental to the intended mathematical structure. The mediation thus provided by a presentation shows how to loosen Benacerraf's dilemma. Group presentations are a good example to consider: $[a, a^n = 1]$ is a presentation of the cyclic group $\mathbf{Z}/n\mathbf{Z}$ among other possible presentations, and yet produces this cycle group itself. The presentation is neither a mere sequence of symbols (as the formalist would have it) because it does present something, nor a proper name (as the platonist would have it) because it has an internal structure that is by itself informative and does not point to any external object. Moreover, tracing back "the" cyclic group with $n$ elements to the presentation $[a, a^n = 1]$ shows how one may (and does indeed) deal with the former on the much more cognitively

tractable basis of the latter. Of course, other presentations of $\mathbf{Z}/n\mathbf{Z}$ are available, in particular presentations that are not specifically presentations by generators and relations. One of the tasks of mathematics is precisely to link all the known presentations together as equivalent presentations of the same "thing."

The "same" mathematical object entertained by multiple presentations is the same object only derivatively, because of the provable equivalence of all these presentations. This kind of identity is a source of increased complexity. Indeed, the criteria that rule the equivalence of two distinct presentations, as well as the logical resources (in the broad sense) available to ascertain it, have varied significantly in history. In particular, when resources are too weak, no connections may be made. The equivalence relating certain geometric entities with algebraic equations was beyond consideration before Descartes' analytic geometry. The proof of the equivalence of mathematical objects in the sense of the existence of categories (such as that between the category of Boolean algebras and the category of Stone spaces) required the development of category theory. The variety of tools to set an equivalence is not only historical, i.e. diachronical. For instance, the equivalence of categories is a global way of conceiving equivalence that differs in that respect from the set-theoretic notion of isomorphism, yet both notions belong to contemporary mathematics and may not be assigned to different periods of its history. Thus, the framework within which the comparison of different presentations may be established in a more or less fine-grained way, and the whole apparatus that Bourbaki described in general terms as "transport of structures," are themselves part of the process from which stable mathematical objects emerge. In other words, the emergence of stabilized ways of assessing the equivalence of (thereby) stable mathematical objects runs concurrently with the emergence of those objects. Of course, it always remains an option to maintain that stable objects come first and that equivalence proofs cannot be integral to them, either because a bundle of proofs can never make up an object, or because equivalence proofs presuppose as objects the very items they show to be equivalent.[15] What is, then, the way out of the difficulties to which platonism leads? A more constructive answer to meet the challenge is this: any equivalence proof bears on two mathematical constructions which, in a sense, already have the status of pre-objects, but a genuine mathematical object stands out only after several equivalence proofs have secured some invariant, e.g., to take a very elementary example, the mathematical way of looking at a triangle begins with disregarding its size or actual position.

There is one last point which should be cleared up. As already mentioned, Kant explains that, in the first antinomy, the idea of the world is too small for the concept of the understanding in the thesis, and too big for it in the antithesis. In the case of Benacerraf's dilemma, one could be tempted to consider things the other way around: that in the equivalent of the thesis (the standard conception), the idea of mathematical objectivity is too big for the understanding, and that in the equivalent of the antithesis (the combinatorial view), it becomes too small. As Benacerraf says:

> [A] typical "standard" account (at least in the case of number theory or set theory) will depict truth conditions in terms of conditions on objects whose nature, as normally conceived, places them beyond the reach of the better understood means of human cognition (e.g., sense perception and the like). […] [P]ostulational stipulation makes no connection between the propositions and their subject matter — stipulation does not provide truth. At best, it limits the class of truth definitions (interpretations) consistent with the stipulations. But that is not enough.
>
> Benacerraf (1973: 667–668, 679)

How should one explain this manifest inversion? As a matter of fact, there is none. The idea of mathematical objectivity, as supplied by the standard conception, is too small indeed, and the same idea, as supplied by the combinatorial view, is undoubtedly too big. This is because the yardstick is not so much our cognitive power as the open-ended series of presentations that mathematicians come up with within the course of history. The standard conception closes the process too early and thus supports too narrow an idea of a mathematical object: the content of a mathematical concept is fixed once for all ("*the* number 3 *is*…") despite the fact that it continues to be enriched by new theorems, so that the problem is now to decide whether one really sticks to the same object when some new presentation of it is put forward. (Think of the parabola when it came to be construed as a scheme, or of the number 3 when it came to be construed as an ordinal set). On the contrary, the combinatorial view contends that a mathematical object is not in any way different from the (thereby too big) complete series of all that is and will be proved about it, which now raises the symmetrical problem of accounting for genuine ruptures in the history of mathematics. Both opponents neglect the deep historical nature of mathematical objectivity—which should come as no surprise since the main proponents on either side (Tarski and Hilbert, respectively) advocated a mainly logical and a-historical view of mathematics.

## 3.4 Conclusion

As a starting point, we saw that an analogy may be drawn between Kant's antinomies and Benacerraf's dilemma: dogmatism assumes the role of the semantic horn and empiricism assumes the role of the epistemic horn. Benacerraf's dilemma can then be reconsidered as a mathematical antinomy about mathematical objectivity. The analogy turned out to be remarkably steady and sharp.

In that perspective, "Mathematical Truth" may be read as an attempt at reconstructing a naïve philosophy of mathematics in order to better overcome it, and to do so in quite the way that Kant undertook to overcome rational cosmology. Two other seminal papers by Benacerraf, Benacerraf (1965) and Benacerraf (1981), confirm that both horns are inconsistent, and that, to resort once more to Kantian terminology, the antinomy is genuinely mathematical, as opposed to dynamical.

The main motivation of the analogy, though, was the prospect of transposing Kant's solution so as to suggest a way of solving Benacerraf's dilemma *proper*. The upshot of the analogy is that a mathematical object can never be given in itself, because it consists in the open-ended series of all its possible presentations, in the sense specified above. As Benacerraf's own example shows, Ernie's and Johnny's "accounts" constitute two different presentations of the natural numbers. The standard conception and the combinatorial view are both wrong because they crystallize an open-ended series either into a closed referent or into a complete infinite series. Both give rise to the dilemma and face further problems of their own, in particular that of accounting for well known historical shifts in the definitions of classical mathematical objects. On the contrary, the solution to Benacerraf's dilemma that one may draw from the analogy with Kant's antinomies calls for a more history-sensitive analysis of mathematical objectivity. It acknowledges that mathematical objects do exist, but proposes to conceive of each as a series of equivalent presentations which remains in the making, thus revisable, i.e. as objects which may not be separated from the proofs explaining that the equivalent presentations it gathers are indeed equivalent, and how they are.

**Notes**

1. Kant [1787] (1998: A426-427/B454-455; 470–471).
2. Kant [1787] (1998: A434-435/B462-463; 476–477).
3. See Kant [1787] (1998: A417/B445; 465).
4. See Kant [1787] (1998: A465-466/B493-494; 498): "In the assertions of the antithesis, one notes a perfect uniformity in their manner of thought and complete unity in their maxims, namely a principle of pure *empiricism*, not only in the explanation of appearances in the world, but also in the dissolution of the transcendantal ideas of the world-whole itself. Against this the assertions of the thesis are grounded not only on empiricism within the series of appearances but also on intellectualistic starting points, and their maxim is to that extent not simple. On the basis of their essential distinguishing mark, however, I will call them the *dogmatism* of pure reason."
5. See Kant [1787] (1998: A466-472/B494-500; 498–501).
6. Admittedly, both opponents of Kant's first antinomy agree to refer to "the world," whereas the formalist, in Benacerraf's dilemma, is unwilling to acknowledge any mathematical referent whatsoever. This seems to weaken the comparison between the empiricist antithesis in Kant and the epistemic horn in Benacerraf. In fact, this should rather urge one to understand how inconsistent the concept of the world is, as the empiricist describes it, namely as being both an *infinite* series and a *given* totality (see Kant [1787] (1998: A418/B445-446; 465)). Whereas Benacerraf's formalist claims not to refer to anything, Kant's empiricist claims to refer to something which actually cannot be anything. This difference qualifies the parallel, but only to a very limited extent, because the

empiricist of the first antinomy does not invoke any individual entity—only an infinite series which, even if it supposed to be completed, can hardly be compared to a referent in the usual sense of the word.

7. In Benacerraf (1965: 71–72), Benacerraf considers shifting back from "the" natural numbers to the space of all ω-progressions, but remarks that "ordinary communication" requires one single distinguished sequence of notations. This is echoed by the semantic horn in Benacerraf (1973): for practical reasons, we need to speak of "the" natural numbers. But we should step back from that habit if we wish to be faithful to the way mathematics really works. Speaking of *the* natural numbers is really just a way of speaking of *any* ω-progression. The structure of natural numbers is an abstraction that results from the desire to single out one privileged ω-progression. From that point of view, as Benacerraf (1965) points out, the main misconception concerning mathematical objects consists in conceiving each of them as being both a proxy for a whole class of particular objects and a particular member of that class.
8. About the notion of presentation, see Sect. 3.3.1.
9. See, for instance, Benacerraf (1965: 58–62).
10. Russell (1903: §329; 349).
11. See also Kant [1787] (1998: A518-523/B546-551; 525–528).
12. Kant adds: "The series of appearances is to be encountered only in the regressive synthesis itself, but is not encountered in itself in appearance, as a thing on its own given prior to every regress."
13. Stanley and Williamson (2001: 427), referring to the Fregean notion of mode of presentation, have transposed it into a practical context. To stick to their example: the ascription to Hannah of a certain knowing how (knowing how to ride a bicycle) is described as a "*practical* mode of presentation of Hannah's propositional knowledge that a particular way of riding a bicycle is a way for her to ride a bicycle" (see op. cit.: 428–429). Whereas a linguistic or semantic mode of presentation designates the particular way in which a proposition is entertained, a practical mode of presentation is a particular way of expressing or understanding "ways of engaging in actions" (*op. cit.*: 436). The analysis of the notion of practical mode of presentation is still to be developed, as the authors concede (*op. cit.*: 429). This paper claims that the notion of *mathematical* presentation is certainly clearer than that of practical mode of presentation, in particular because the former does not rely on a mere parallel with the Fregean one, whereas the latter does; it does indeed correspond to one of the main cases that Frege had in mind.
14. Taking presentations seriously, as essential devices in between mathematical structures and the empirical systems that instantiate them, would be a way to overcome the "identity problem" that has been raised against *ante rem* structuralism (see, in particular, Keränen 2006; Shapiro 2006). The identity problem comes from the existence of non-trivial automorphisms in certain structures (such as conjugacy in the field of complex numbers). The solution that could be given to that problem is that a non-trivial automorphism of a given structure is always, in fact, an isomorphism between two presentations associated with that

structure (for instance, conjugacy is an isomorphism between $\langle \mathrm{C}, i, -i \rangle$ and $\langle \mathrm{C}, -i, i \rangle$, taken as presentations of the same structure). This solution requires to admit that a structure is never accessible outside some of its presentations, which clearly qualifies realist structuralism, without giving in, however, to the empiricist thesis that one can have access only to systems (i.e. to concrete instances of structures).

15. This objection was raised by Marco Panza and Achille Varzi in two different ways. I thank them both for their remarks.

## References

Benacerraf, P. (1965). What numbers could not be. *The Philosophical Review, 74*(1), 47–73.

Benacerraf, P. (1973). Mathematical truth. *The Journal of Philosophy, 70*(19), 661–679.

Benacerraf, P. (1981) Frege: The last logicist. *Midwest Studies in Philosophy, 6*(1), 17–36.

Brentano, F. C. (1874). *Psychologie vom empirischen Standpunkt*. Leipzig: Verlag von Duncker & Humblot.

Frege, G. [1892] (1952). On sense and reference. *Translations from the philosophical writings of Gottlob Frege* (P. Geach & M. Black, Eds., M. Black, Trans., pp. 56–78). Oxford: Basil Blackwell.

Kant, I. [1787] (1998). *Critique of pure reason* (P. Guyer & A. W. Good, Eds., Trans.). [from the second edition of *Kritik der reinen Vernunft*, Johann Friedrich Hartknoch, Riga, 1787]. Cambridge: Cambridge UP.

Keränen, J. (2006). The identity problem for realist structuralism II. A reply to Shapiro (Chapter 6). In F. MacBride (Ed.), *Identity and modality. New essays in metaphysics* (pp. 146–163). Oxford: Oxford UP.

Russell, B. (1903). *The principles of mathematics*. Cambridge: Cambridge UP.

Shapiro, S. (2006). Structure and identity (Chapter 5). In F. MacBride (Ed.), *Identity and modality. New essays in metaphysics* (pp. 109–145). Oxford: Oxford UP.

Stanley, J., & Williamson, T. (2001). Knowing how. *The Journal of Philosophy, 98*(8), 411–444.

von Dyck, W. F. (1882). Gruppentheoretische Studien. *Mathematische Annalen, 20*(1), 1–44.

# Chapter 4
# On Benacerraf's Dilemma, Again

**Marco Panza**

In spite of its enormous influence, Benacerraf's dilemma admits no standard unanimously accepted formulation. This mainly depends on Benacerraf's having originally presented it in a quite colloquial way, by avoiding any compact, somehow codified, but purportedly comprehensive formulation (Benacerraf 1973). But it also depends on Benacerraf's appealing, while expounding the dilemma, to so many conceptual ingredients so as to spontaneously generate the feeling that most of them might in fact be inessential to it. Apart from the almost unanimous agreement on the fact that, despite Benacerraf's appeal to a causal conception of knowledge throughout his exposition, the dilemma does not rely on it, there is still no agreement about which of these many ingredients is essential and which should be left aside, in agreement with an Ockhamist policy, so as to obtain a minimal version of the dilemma.

I will firstly offer a discussion of this matter (Sect. 4.1), with a particular attention to Field's reformulation of the problem (especially in Field 1989), in order to identify two converging and fundamental challenges addressed by Benacerraf's dilemma, respectively to a platonist and to a combinatorialist philosophy of mathematics, in Benacerraf's own sense of these terms (Sects. 4.2 and 4.3, respectively). What I mean by saying that these challenges are convergent is that they share a common core which

I thank for valuable comments and suggestions Paul Benacerraf, Stefan Bujisman, Annalisa Coliva, Fabrice Pataut, Andrea Sereni, Göran Sundholm, and Gabriele Usberti.

M. Panza (✉)
CNRS, IHPST (CNRS and University of Paris 1 Panthéon-Sorbonne),
13, rue Du Four, Paris 75006, France
e-mail: marco.panza@univ-paris1.fr

and

Chapman University, One University Drive, Orange, CA 92866, USA
e-mail: panza@chapman.edu



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_4

embeds a crucial puzzle for any plausible philosophy of mathematics,[1] and that they suggest a way out along similar lines. Roughing these lines out is the purpose of the last two sections of the paper (Sects. 4.4 and 4.5).

## 4.1 Field's Reformulation of Benacerraf's Challenge to a Platonist: Is the Problem Really About Truth and Knowledge?

Unquestionably, Benacerraf's purpose is to keep the reader's attention focused on an alleged contrast between two kinds of philosophical concerns about mathematics, one doubtlessly epistemological, the other apparently semantic, though possibly ontological in nature. Having been originally written for a presentation at a symposium on mathematical truth, his paper mentions this notion in its very title and, from its very first lines, declares its interest for mathematical knowledge, explicitly presented as a notion depending "on how truth in mathematics is properly explained" (Benacerraf 1973: 661). Still, under some readings, the dilemma appears to be eventually independent of both the notions of (mathematical) truth and (mathematical) knowledge, and to be concerned with the connection between mathematical beliefs, or possibly their formation process or justification, and the subject matter of mathematics, i.e. whatever mathematics is taken to be about, provided that it is taken to be about something.

This, at least, is what is suggested by Field's reformulation of the problem, which ascribes to it the form of a challenge to 'mathematical realism' or "platonism", conceived as "the view that there are mathematical entities and that they are in no way mind-dependent or language-dependent" (Field 1989: 228), that they "bear no spatio-temporal relation to us [, and] […] do not undergo any physical interactions […] with us or anything we can observe" (Field 1989: 27), in short that they are both mind- and language-independent, and abstract.[2]

One of Field's explicit purposes is precisely to adapt the challenge to this characterization of platonism, which is, as such, independent of any appeal to truth and knowledge. According to this picture, a mathematical platonist is not required to take mathematical statements to be true in some sense of 'true' "more loaded" that a mere "disquotational" sense (Field 1989: 228–229; see also Field 1988: 62–63). The platonist is merely required to maintain that "his or her own states of mathematical beliefs, and those of most members of the mathematical community […] are highly correlated with the mathematical facts" (Field 1989: 230; see also Field 1988: 62), namely the "facts about [the] mathematical entities" (Field 1989: 232) that he or she takes to obtain. Once this is admitted, the challenge becomes independent, as such, not only of any appeal to some non-merely-disquotational conception of truth, but also of "any theory of knowledge", and then, of "any assumption about necessary and sufficient condition for knowledge" (Field 1989: 232–233). In short, it "can be put without use of the term of art 'knows' and '[…] without talk of truth'" (Field 1989: 230; see also Field 1988: 62).

What Field means when he claims that, for a platonist, the states of mathematical beliefs of most members of the mathematical community are highly correlated with the mathematical facts is that "for most mathematical sentences that you substitute for '*p*', the following holds: [i]f mathematicians accept '*p*' then *p*" (Field 1989: 230; Field 1988: 62; see also Field 1989: 26). The challenge consists, then, in requiring of a platonist an appropriate explanation of such a "systematic correlation" or "general regularity" (Field 1989: 231), "an explanation of how it can have come about that mathematicians's belief states and utterances so well reflect the mathematical facts" (Field 1989: 230; see also Field 1988: 62). It seems plain that accepting *p* is here taken to be the same thing as believing that *p*, and that mathematicians' utterances are taken to be content-transparent expressions of mathematicians' belief states. This suggests the following rephrasing of the challenge: how can a platonist explain that, at least in the great majority of cases, a mathematician has the mathematical belief that *p* only if it is a (mathematical) fact that *p*?

From Field's own perspective, the question is rhetorical, of course, since according to him "there seems *prima facie* to be a difficulty in principle in explaining the regularity" (Field 1989: 230–231). Later on in his paper, this *prima facie* difficulty becomes a principled impossibility, and the challenge turns into a negative dictum:

> […] we should view with suspicion any claim to know facts about a certain domain [namely mathematics, in the case at issue] if we believe it is impossible in principle to explain the reliability of our beliefs about the domain.
>
> Field (1989: 230); see also Field (1989: 26)

Two things might seem eccentric about this way of setting the debate: it relies on the quite controversial and loaded epistemological notion of reliability, which the previous considerations do not take into account, apparently; it appeals to the notion of mathematical knowledge, which seems to contradict Field's initial proposal.

A way to answer the former worry is to observe that Field is not here resorting to the reliability of some sort of justification or belief formation process, but rather to the reliability of the relevant beliefs themselves. Of course, one might stipulate that a belief is reliable just in case its formation process is reliable. But this does not seem to be what Field is driving at. He seems to take the reliability of mathematical beliefs to be the same thing as their reflecting the mathematical facts, in the sense of the systematic correlation or general regularity previously mentioned.[3]

A way to answer the latter worry is to observe that mathematical knowledge only enters the matter *a fortiori*, so to say, and independently of any specific view about it: what Field seems to be saying is that, if this correlation or regularity is not explained, there is no room for a platonist to provide an appropriate account of mathematical knowledge, whatever his or her conception of knowledge might be.[4]

The challenge does not seem, then, to be different in nature; the crucial point is the same as before: how can a platonist explain that, in the great majority of cases, a mathematician has the mathematical belief that *p* only if *p*?

But—one could object—is a platonist actually required to maintain that, in the great majority of cases, a mathematician has the mathematical belief that *p* only if *p*?

Even if it is admitted that the reference of the term 'mathematician' is fixed with some degree of certainty, the answer still largely depends on what is meant by 'in the great majority of cases' and 'mathematical belief.' If one takes a mathematical belief to be any belief that can be expressed by means of a statement using a mathematical vocabulary (or a vocabulary widely recognized as being a mathematical one), or even using only such a vocabulary (together with the appropriate logical constants), and the majority of cases to be the majority of mathematical beliefs that mathematicians have now, had in the past, or will have in the future, the answer seems to be negative. Leaving logically possible cases of collective hypnosis or hallucination on the side, it remains that many mathematicians hold, at some point in time, opposite beliefs since, presumably, many of these beliefs are not grounded on proofs, or at least on any widely accepted proof, but rather depend on methodological, philosophical, aesthetic, or even mystical attitudes or convictions.[5] In other words, much of what mathematicians believe about what they take to be the subject-matter of mathematics, even of pure mathematics, is open to controversy within the mathematical community itself. There is, then, no reason to think that someone who considers that there are mathematical objects, and that they are mind- and language-independent and abstract should also maintain that, in the great majority of cases, a mathematician has the mathematical belief that *p* only if *p*.

Field's point seems on the contrary much more plausible if the range of mathematical beliefs is restricted to beliefs somehow secured within purely mathematical theories that are widely accepted by the mathematical community. For short, call these beliefs 'mathematical theory-tied beliefs'.[6]

That the challenge is as a matter of fact restricted to these beliefs is something that Field himself suggests. He remarks that "as mathematics has become more and more deductively systematized, the truth [disquotationally understood, I suppose] of mathematics has become reduced to the truth of a smaller and smaller set of basic axioms", with the result that what a platonist needs to explain is only the alleged (by him or her) circumstance that "for all (or most) sentences '*p*' […] [,] if most mathematicians accept '*p*' as an axiom, then *p*", or better, that "either *p*, or [most] mathematicians don't take '*p*' as an axiom" (Field 1989: 231). So conceived, the challenge is echoed by Heck who claims, rephrasing Benacerraf's dilemma:

> […] we lack […] an explanation of how we come to know the axioms, be these the axioms of some developed mathematical theory or those propositions which are, in a less developed theory's present state, typically assumed without proof. More precisely, it is not obvious why there should be any relation at all between our belief that the axioms are true and the facts of mathematics as the platonist conceives them, why our beliefs should reliably reflect how things stand with the sets or the numbers, or whatever.
>
> Heck (2000: 128)

Field and Heck clearly consider that it is easy to meet the challenge for whatever mathematical theory-tied belief in case it is also met for the axioms of the relevant theories. Still, this is so only if these theories consist in formal axiomatic systems and if these systems are sound in the appropriate sense, i.e. if they are such that, if their

axioms reflect the mathematical facts, then so do their theorems. Even if it were taken for granted that the former condition is met, one would still have to argue that the latter is met as well: one would still have to explain how it happens that the deductive rules of the relevant theories fit in with the way mathematical facts are related to each other. For this reason, and also because one would have to take into account other kinds of theories, either formal or not, that do not reduce to axiomatic systems, it would be more appropriate to generalize what Field and Heck say on mathematical axioms to any sort of liminal assumptions of the relevant theories, certainly including axioms, but also other kinds of stipulations or presumptions (either explicitly governing formal deductions, like deductive rules, or playing a role in informal but widely accepted proofs).

Still, if it is admitted that meeting the challenge for these liminal assumptions makes it be automatically meet (or, at least, makes easy to meet it) for any mathematical theory-tied beliefs are concerned, why should one restrict its formulation to the former? This does not make it easier to meet. For the sake of all-inclusiveness, it is therefore advisable to state the challenge as applying to mathematical theory-tied beliefs in general.

Finally, if the content of these beliefs is taken to be stable under the variety of cognitive subjects that hold them and of the cognitive contexts in which they do (as Field and Heck seem to admit), the challenge may be stated without any appeal to mathematicians as bearers of these beliefs: all what is relevant are the beliefs themselves.[7]

In the end, the question seems to be the following, then: how can a platonist explain that mathematical theory-tied beliefs reflect the mathematical facts (in the sense specified above)?

It remains, however, that not appealing to mathematicians as the bearers of the relevant beliefs when stating the challenge is not the same as taking it to be independent of what mathematicians do, namely of their providing justifications for these beliefs. By definition, a belief is mathematical theory-tied just in case it comes together with a consensual epistemic practice to secure it: typically a generally accepted justification for it, or at least a consensual admission that it has an acceptable justification, namely a widely accepted proof or another sort of direct or indirect ground, such as those one usually appeals to when supporting mathematical axioms or other liminal assumptions of mathematical theories.

Hence, although neither Field's nor Heck's formulation of the challenge appeals to the justification of the relevant beliefs (either under the form of a proof or of any sort of suitable ground), and despite Field's claim that "Benacerraf's challenge […] is not […] a challenge to our ability to justify our mathematical beliefs, but […] a challenge to our ability to explain the reliability of these beliefs" (Field 1989: 25), the justification for the relevant beliefs seems to be an indispensable ingredient of it.[8] Field's insistence on the idea of a systematic correlation, or of a general regularity, suggests, then, that the required explanation could be offered only insofar as a stable connection (regularly and/or systematically operating under the variation of *p*) between the mathematical fact that *p* and the consensually accepted justifications for the belief that *p* may be identified. Taking the identification of this connection to be

an essential ingredient of a theory of mathematical knowledge, or preferring to avoid any appeal to so loaded a term as 'knowledge' for describing the problematic setting at issue, depends on a terminological choice rather than on some substantial option. What matters is that, according to Field, a platonist could hardly be credited with a decent epistemology (broadly understood as an analysis of the virtues that mathematics has for us) if he or she were not able to answer this challenge.[9]

## 4.2 Another Way to Understand Benacerraf's Challenge to a Platonist

When Benacerraf's dilemma is seen in this light, its original formulation seems to depend on the requirement that the connection to be explained (which in the original settings takes the form of a "connection between the truth conditions of *p* […] and the grounds on which *p* is said to be known": Benacerraf 1973, p. 672) hinge on a "causal relation" (*ibid*., p. 671) between the epistemic subjects having the relevant beliefs and the constituents of the relevant facts, namely the objects that these facts are supposed to be about[10]: a relation allowing one to account for the justification of these beliefs by admitting that these subjects are causally affected by these objects, and that the justification just results from this.

Not only is it worth our while to give up this requirement, since, as famously observed by Hart (1977: 125–126), "superficial worries about the intellectual hygiene of causal theories of knowledge are irrelevant […] and misleading […], the problem [being] not so much about causality as about the very possibility of natural knowledge of abstract objects". This is also indispensable if we are to avoid begging the question by putting too heavy a burden on the platonist—admitting, in agreement with Field's account, that mathematical objects are mind- and language-independent, abstract, and that mathematical facts are facts about these mind- and language-independent *abstracta*.

Moreover, Sereni's argumentation in the paper included in the present volume suggests that any plausible requirement concerning the nature of the relevant connection is likely to either beg the question or yield too unspecific a challenge.

Should we conclude with Sereni that the challenge Benacerraf's dilemma addresses to the platonist, even if understood in the minimal form I have described above along Field's lines, is either ill-posed or unspecific?[11]

I do not think so. I think on the contrary that for the challenge to be correctly formulated and specific,[12] it is enough to stay away from any requirement on the nature of the connection between the mathematical theory-tied belief that *p* and the (mathematical) fact that *p*. All what is to be required is that the justification of the former (i.e. the consensual epistemic practice that secures the belief and makes it a mathematical theory-tied belief)[13] be a justification that the latter obtains.

To make this point clear, let us reflect once again on Field's picture. Not only does this picture depend on the characterization of the platonist as someone who

maintains that there are mathematical objects, and that they are mind- and language-independent and abstract, but it also depends on the claim that, according to the platonist, the relevant mathematical beliefs have a propositional content, and that this content is that a mathematical fact, namely a fact about these very objects (presumably consisting in their being in certain relations with one another), obtains. In other terms, this picture seems to take for granted that a (mathematical theory-tied) belief that *p* is just the belief that the (mathematical) fact that *p* obtains, i.e. that some appropriate (mathematical) objects are related to one another in a certain way.

This is quite natural to admit from a platonist perspective and does not seem as such to require any further explanation. What is by far much less natural to admit and requires indeed an explanation is that a justification for a mathematical theory-tied belief that *p* be a justification that the mathematical fact that *p* obtains, that is, a justification that the relevant mathematical objects are related to one another in a certain way. Taking a justification of a mathematical theory-tied belief to meet this requirement is in itself a very strong assumption: an assumption that a platonist cannot merely take for granted and that must indeed be argued for.

Here is, in my view, the basic challenge that Benacerraf's dilemma addresses to the platonist: the platonist is required to explain how the justification of a mathematical theory-tied belief (i.e. either an argument supporting an axiom or some other kind of liminal assumption of a mathematical theory, or a proof within such a theory) may count as the justification that a mathematical fact, conceived as a fact about the relevant mathematical objects, obtains. The crucial question is not concerned with what, from a platonistic point of view, could guarantee that the relevant mathematical beliefs reflect (in Field's sense) the facts about the mathematical objects, but rather with the very possibility of taking the justifications for these beliefs to be justifications that such facts obtain.

Consider the belief that 5 + 7 = 12. It should be plain that this belief is both mathematical and (as opposed, for instance, to the belief that 5 + 7 = 13) theory-tied. One might have many different ideas about what should count as a justification for it, but it should be clear that most of us would spontaneously consider that a proof that 5 + 7 = 12, or of '5 + 7 = 12' within some accepted version of arithmetic, constitutes indeed such a justification (and even a suitable and reliable one if it were admitted that this belief could also have unsuitable or unreliable justifications). Now, there is no doubt that such a proof justifies a belief. The point is whether a platonist can, in agreement with the spontaneous pronouncement of most of us, take this belief to be the very belief that 5 + 7 = 12.

According to Field's construal of platonism, this last belief is the belief that: (i) there are numbers, namely the numbers 5, 7, and 12; (ii) they are mind- and language-independent abstract objects; (iii) they are related to one another by the additive relation expressed by the statement '5 + 7 = 12.' Hence, if this picture is

admitted, the point is whether a platonist can take a proof that 5 + 7 = 12, or a proof of '5 + 7 = 12' within some accepted version of arithmetic, to provide a justification that facts (i)–(iii) obtain, or at least—supposing that it has been previously justified on independent grounds that facts (i)–(ii) obtain—that fact (iii) also obtains, rather than to provide a justification that it is a theorem of some version of arithmetic that 5 + 7 = 12, or that '5 + 7 = 12' is such a theorem.

To avoid additional worries, let us focus for the sake of argument on categorical versions of arithmetic, say on PA2. Following a platonist, let us also take for granted —still for the sake of argument—that the singular terms of PA2, or of whatever categorical version of arithmetic, have the same reference as the corresponding numerical terms we use, both in our ordinary informal arithmetic and in our everyday parlance (provided these terms have any reference at all). It would be a crucial challenge for a platonist, as described by Field, to explain how it can happen that a proof within such a suitable accepted version of arithmetic justifies that the facts that (i)–(iii) obtain or, alternatively (under the mentioned condition), that the fact that (iii) obtains.

A similar concern also applies to other sorts of mathematical theory-tied beliefs, namely those pertaining to the axioms or other liminal assumptions of mathematical theories. Take as another example the axiom of existence and unicity of the successor for natural numbers in one of its (formal or informal) formulations, which, still for the sake of argument, we shall take as being all about the same objects (provided they are about objects at all). According to Field's picture, for a platonist to believe this axiom is to believe that: (i) there are natural numbers; (ii) they are mind- and language-independent abstract objects; (iii) every such number has a single successor. Now, many grounds have been offered for this axiom, and all of them undoubtedly justify a belief. What is far from clear is whether these grounds are justifications that facts that (i)–(iii) obtain, or, at least—supposing that it has been previously justified on independent grounds, that facts (i)–(ii) obtain—that fact (iii) also obtains, rather than justifications that this axiom is a suitable axiom for a suitable version of arithmetic. Again, an obvious challenge for a platonist, as described by Field, is to explain how it can happen that the former option holds rather than the latter.

The two cases are indeed not equivalent. In the former, the justification is regimented so as to make any doubt about its internal correctness or accuracy immaterial. In the latter, the justification is either essentially informal, in which case it is open to doubts or plausible scepticism, or it is regimented within a metatheory (as it happens for justifications based on proofs of completeness or categoricity), in which case it openly conforms only to the second of the two options considered, since a proof within a metatheory possibly justifies that some facts about the relevant theory obtain, but certainly not that some facts about whatever the theory is about obtain as well. Still, this difference does not seem to affect the point I want to make, since this point is not concerned with the accuracy or perfection of the justification, but rather with what it is a justification of, and stands on its own feet even if it is granted for the sake of argument that our formal and informal mathematical argumentations are all about the same objects (provided, once again, that

they are about objects at all). I therefore take the challenge to be the same in the two cases and to apply quite generally to any mathematical theory-tied belief: it is the challenge to explain how a justification for such a belief that $p$ can be a justification that the fact that $p$ obtains.[14]

When it is couched in these terms, the challenge seems to be close to the one to which Burgess and Rosen have reduced Benacerraf's dilemma: "granted that belief in some theory is justified by scientific standards, is belief in the truth of that theory justified?" (Burgess and Rosen 1997: 48). This reduction goes through many intermediary stages that I shall account for here only partially. (One will find a compendious account and a stringent criticism of it in Hale 1998: 162–163). Burgess and Rosen begin with their own construal of Field's challenge: if it is true that "when mathematicians believe a claim about mathematicalia, then that claim is true", an explanation of this is in order (Burgess and Rosen op. cit.: 41–42). They then admit, with Field, that the explanation should not be required to concern "what follows logically or analytically from what" and restrict the issue to "axiomatic beliefs, ones that are not believed simply because they follow logically or analytically from other, more basic beliefs" (Burgess and Rosen op. cit.: 45). Next, they appeal to the possibility of rephrasing all mathematics within set theory (and to other connected considerations that we may leave here on the side) in order to further restrict the challenge to the following: "granted that belief in standard set theory is justified by scientific standards, is belief in the truth of standard set theory justified?" (Burgess and Rosen op. cit.: 47). Finally, they argue, in agreement with what Benacerraf and Putnam contend in the introduction to Benacerraf and Putnam (1983), at page 35, that the problem is not "peculiar to mathematics" (Burgess and Rosen loc. cit.),[15] and remark, in agreement with Field, that truth is here to be intended disquotationally, so as to reach the last reformulation quoted above at the very last step.

This long detour already suggests, however, that the way in which Burgess. and Rosen consider their own version of the challenge is quite different from the way I consider mine. This difference is blatant in the following comment:

> Once it is put in this last form, it becomes clear that the […] challenge presupposes a 'heavy duty' notion of 'justification' — one not just constituted by ordinary commonsense standards of justification and their scientific refinements […]. To put the matter another way, once it is put in this last form, it becomes clear that the question or challenge is essentially just a demand for a philosophical 'foundation' for common sense and science — one that would show it to be something more that just a convenient way for creatures with capacities like ours to organize their experience — of the kind that Quine's naturalized epistemology rejects.
>
> Burgess and Rosen (1997: 48)

It is clear, then, that Burgess and Rosen construe the challenge as a request for an ultimate legitimization of our theories, both mathematical and scientific, depending on what there is and on how things actually are, a request that openly violates naturalism (see, e.g., Liggins 2010: 73).

Although I do not see any reason for a philosophy of mathematics to abide by naturalism, I do not construe the challenge in this way. I understand it as the

requirement of an account of mathematics able to explain how mathematical theories, which we devise and select according to our own standards, can say what they say according to a platonistic construal, and how, from then on, the justifications which are germane to mathematics may also be justifications for mathematical beliefs platonistically construed.[16] I am not asking for anything more when I argue that an explanation of how a justification of the mathematical theory-tied belief that *p* may also be a justification of the fact that *p* obtains must be given.[17]

I fail to see why the challenge, thus understood, should beg the question for platonism, as construed by Field. If it did, the begging would come from the requirement that a justification for a mathematical theory-tied belief not be a justification for something distinct from what the platonist takes that belief to be. In other words, what would beg the question for such a platonist would be the very request to offer an epistemology of mathematics that should be both plausible and compatible with the platonist view of what mathematics is (about).

When understood in this way, the challenge that Benacerraf's dilemma addresses to the platonist is neither concerned with the reliability of the relevant justifications, nor with any other possible condition that these justifications could be required to meet (beside that of being indeed justifications for the relevant beliefs).[18] This cuts short a number of questions, arguments and counter-arguments that are often evoked in discussions allegedly concerned with Benacerraf's dilemma. Let us then leave these questions, arguments and counterarguments on the side and ask now which challenge, if any, the dilemma offers to a combinatorialist (in Benacerraf's sense of that word).

## 4.3 Benacerraf's Challenge to a Combinatorialist

A simple way to answer the former challenge is to deny that the mathematical fact that *p* is distinct from the fact that it is a theorem or a liminal assumption of an accepted mathematical theory that *p*. Insofar as it seems quite implausible (since contrary to the evidence coming from mathematical practice) that the former fact reduces to the latter, this depends on admitting that there is nothing like a mathematical fact that *p* other than the mere fact that it is a theorem or a liminal assumption of an accepted mathematical theory that *p*. For example, there is nothing such as the mathematical fact that $5 + 7 = 12$, or that every natural number has one and only one successor over and above, respectively, the mere fact that it is a theorem of an accepted mathematical theory that $5 + 7 = 12$, and that it is a suitable axiom for a suitable version of arithmetic that every natural number has only one and only successor.

Field denies that this solution is open to a platonist on the ground that it is incompatible with the view that there are mathematical objects and that they are mind- and language-independent and abstract; unless, of course, the platonist is ready to admit that mathematical facts are not facts about these objects, which

would then make his view quite immaterial as a basis for a philosophical account of mathematics.

It is, instead, quite in line with the views about mathematical truth that Benacerraf calls 'combinatorial', according to which "the truth conditions for arithmetical [but one could in general say 'mathematical'] sentences are given as their [...] derivability from specified sets of axioms", provided that such a derivability is broadly construed, or that the requirement of completeness (understood as the requirement that a truth-value be assigned to each statement of the language of the relevant theorem) is abandoned, so as to avoid the difficulties brought forth by Gödel's incompleteness theorem (Benacerraf 1973: 665). According to Benacerraf, the "leading idea" behind these views is indeed "that of assigning truth-values to arithmetical [but, once again, one could in general say 'mathematical'] sentences on the basis of certain (usually proof-theoretic) syntactic facts about them" (Benacerraf loc. cit.).

The claim that the there is no mathematical fact that $p$ beside the fact that it is a theorem or a liminal assumption of an accepted mathematical theory that $p$ does not rely on a non-disquotational notion of mathematical truth. This claim thus turns out to be a natural way of expressing the combinatorial views in a weakened setting where any such notion is dismissed, such as the setting of Field's reformulation of the challenge that Benacerraf's dilemma addresses to the platonist. This suggests that the challenge to the combinatorialist could be restated within such a weakened setting as a challenge to a supporter of such a claim.

Although Benacerraf does not insist on this point in his original paper, he clearly remarks that combinatorial views are in need not only of a combinatorial account of mathematical truth, but also of a "new *theory of truth theories*" capable of relating combinatorial truth for mathematics to usual truth for "referential (quantificational) languages" (Benacerraf op. cit.: 669). The challenge he addresses to the combinatorialist seems, then, that of providing such a new theory of truth theories.

It follows, then, that if the notion of truth, for any kind of language, is appropriately weakened, or indeed removed from the setting, one may quite naturally rephrase the challenge in the following way: a combinatorialist must explain how an analysis of a mathematical theory-tied belief that $p$, according to which the content of this belief is that the fact obtains that it is a theorem or a liminal assumption of an accepted mathematical theory that $p$, is related to an analysis of a non-mathematical belief that $q$, somehow connected with the mathematical theory-tied belief that $p$ (for example the belief that '$p$' results from the axioms of the relevant theory by appropriate transformations licensed by the deductive rules of this same theory),[19] according to which the content of this belief is just that the fact that $q$ obtains.

It seems to me, however, that there is more to be said on this matter. Even if such an account were indeed available or, better, before one may even hope to provide it, a combinatorialist should explain how the content of a mathematical theory-tied belief that $p$ is to be precisely determined. Should this content be taken to be that (i) the fact obtains that it is a theorem or a liminal assumption of a specified mathematical theory that $p$, or that (ii) the fact obtains that there is an accepted mathematical theory of which it is a theorem or a liminal assumption that $p$, or that (iii) the fact obtains that for any accepted mathematical theory pertaining to an

appropriate specified branch of mathematics, it is a theorem or a liminal assumption of the theory that *p*?

Moreover, for each of these options, a combinatorialist must also explain what is to be taken as an appropriate justification for the relevant belief, and possibly give that explanation so as to guarantee that our customary and natural conception of what it means for a subject to have the justified belief that *p* be preserved. For example, a proof that 5 + 7 = 12 within PA2 would certainly be a justification of the belief that 5 + 7 = 12 if this belief were analyzed in agreement with (i) and the mathematical theory were identified with PA2, or in agreement with (ii) and PA2 were included in the relevant domain of accepted mathematical theories. But it certainly wouldn't be enough to justify this belief if it were analyzed in agreement with (iii). It seems, then, that a combinatorial view about the content of mathematical theory-tied beliefs and their justification would be plausible only if this view allowed us to specify the content of these beliefs, and of what would count as a justification for them, in a way that would not be either utterly complex, or quite unfaithful to our customary and natural conceptions of both the content and the justification.

I shall not explore the matter further: a better explanation of the difficulties that a combinatorialist should overcome in order to meet the challenge would require more space than allotted here. The point here is to make clear how one should understand the challenge that the Benacerraf's dilemma addresses to a combinatorialist when that challenge is required to dispense with the twin notions of truth and knowledge.

## 4.4 Meeting Both Challenges Simultaneously

One might think that the essential difficulty of the challenge to a platonist is that of filling the gap between the mind- and language-independent abstract objects that the platonist takes to exist, and the human justifications embedded in our mathematical practice. If this were the case, one could try to meet the challenge either by advocating a characterization of platonism alternative to Field's, or by weakening the platonist contention.

According to Field's picture, for a mathematical platonist: (i) mathematics is about appropriately specified objects and non-reducible actual facts concerning them[20]; (ii) these objects are mind- and language-independent; (iii) they are abstract. The more natural option would be to stick to (i) and to give up (ii) and (iii). Call, then, 'minimalist platonist (about mathematics)' anyone who endorses (i), but remains agnostic with respect to (ii) and (iii), i.e. open to either an endorsement or a rejection of either position.[21]

Endorsing (i) entails maintaining that there are mathematical objects.[22] If one endorsed this thesis, but rejected (ii), i.e. argued that there are mathematical objects, but admitted that they could be mind- and language-dependent (and therefore

abstract),[23] one would allow the existence of these objects to be open to conceptualizations that are different from the one according to which existence is a primitive intrinsic condition not open to specification. A minimalist platonist should, then, concede this possibility as well.

According to a minimalist platonist, the content of a mathematical theory-tied belief would still be, however, that a non-reducible fact about mathematical objects obtains, and a justification of such a belief would accordingly still be a justification that this very fact obtains.

One could reply, either that such a minimalist platonist is no longer a genuine platonist one, or that a minimalist platonist fails to have at his or her disposal enough explanatory power to account for mathematical ontology and semantics. It nevertheless seems clear that taking the belief that 5 + 7 = 12 to be a belief about three distinct objects, namely 5, 7, and 12, and the belief that every natural number has a single successor to be a belief about a domain of distinct objects, namely natural numbers, and admitting as well that the fact obtains that 5 + 7 = 12 and that the fact obtains that every natural number has a single successor, and that such facts are non-reducible, amounts to the adoption of a strong philosophical view endowed with a respectable explanatory power, both for mathematical ontology and mathematical semantics, one which is quite different from other views often defended in an anti-platonist perspective. (Note that the view implies that the statements '5 + 7 = 12' and 'every natural number has a single successor' have a semantic structure that parallels their superficial syntactical form, and are true, at least disquotationally.)

Hence, the problem with minimal platonism seems to depend neither on its being too weak or of falling short of explanatory power, nor on its openly being not quite platonist in spirit. It would depend on its not being specifiable without eventually admitting that mathematical objects are, after all, mind- and language-independent, which amounts to either endorsing platonism as described by Field, or to embracing the quite unlikely view that these objects are somehow concrete.

Still, if it were clear that adopting minimal platonism would make it easy to meet the challenge addressed to platonism by Benacerraf's dilemma, both under Field's construal and mine, one could give it a try, even though one would then make a new challenge arises, consisting in the demand for an appropriate specification of this minimal view in order to avoid coming back to an endorsement of (ii). Unfortunately, merely leaving open the possibility of taking mathematical objects to be in some way or other mind- and language-dependent, does not provide an easy way for a plausible explanation of how a justification of the mathematical theory-tied belief that *p* can also be a justification that the fact that *p* obtains, if this fact is conceived as a non-reducible fact about mathematical objects.

The adjective 'plausible' is crucial here. It is intended to mean that the required explanation shouldn't, *mutatis mutandis*, run into the same difficulties as those the combinatorialist views runs into face to Benacerraf's challenge. This would happen indeed if one hoped to provide the required explanation by allowing that mathematical objects be nothing but the items fixed within our current mathematical theories.

Broadly speaking, this is the position defended, albeit in quite different ways, by Shapiro's and Resnik's *ante rem* structuralism (Shapiro 1997; Resnik 1997) and by Linsky and Zalta's version of platonism framed within Zalta's object theory (Linsky and Zalta 1995, 2006; Zalta 1999, 2000). The problem with this position is that it is hardly compatible with meeting the challenge that Benacerraf's dilemma addresses to platonism without specifying what is a fact about mathematical objects, what is the content of a mathematical theory-tied belief and what counts as a justification of such a belief, in a way that is not unreservedly complex, or quite unfaithful to our customary and natural conceptions of what such a fact, content and justification are.

Suppose one argued that, insofar as the fact that 5 + 7 = 12 is a fact about items fixed within a certain version of arithmetic and there called '5', '7' and '12', this very fact is nothing but the fact that a proof within this version of arithmetic ends with '5 + 7 = 12' and analogously, that, insofar as the fact that every natural number has a single successor is a fact about the items fixed within a certain version of arithmetic and and there called 'natural numbers', this very fact is nothing but the fact that this version of arithmetic includes an axiom, or possibly a theorem, just asserting, in the relevant language, that every natural number has a single successor. Given the way the relevant mathematical objects are identified, one could argue that these are non-reducible facts about them—and, from there, that the semantic structure of '5 + 7 = 12' and 'every natural number has a single successor' parallels the superficial syntactical form of these statements. But it would be harder to admit that this way of specifying these facts is faithful to our customary and natural conceptions of what the non-reducible facts about the numbers 5, 7, and 12, and the natural numbers in general are, since, according to these conceptions, these facts are not concerned with any particular theory, but merely with the numbers 5, 7, and 12, and with the natural numbers as such.

The same could be said if it were argued that the fact that 5 + 7 = 12 is nothing but the fact that the items fixed within a certain version of arithmetic (and there called respectively, '5', '7' and '12') stand to each other in such a way that the value of the addition-function fixed within this same version of arithmetic for the first two items taken as arguments, is just the third, and, analogously, that the fact that every natural number has a single successor is nothing but the fact that the items fixed within a certain version of arithmetic (and there called 'natural numbers') stand to each other in such a way that the successor relation thereby fixed within this same version of arithmetic is functional and total.

Things wouldn't be better if one set out to overcome the difficulty by taking the foregoing facts to be nothing but the facts that, for any accepted version A of arithmetic, $5_{\mathrm{A}} + {}_{\mathrm{A}}7_{\mathrm{A}} = 12_{\mathrm{A}}$ (or even $5_{\mathrm{A}} + {}_{\mathrm{A}}7_{\mathrm{A}} = {}_{\mathrm{A}}12_{\mathrm{A}}$), and $\forall x[NN_{\mathrm{A}}(x) \Rightarrow \exists!y[NN_{\mathrm{A}}(y) \wedge SUC_{\mathrm{A}}(x, y)]]$ (or even that $\forall_{\mathrm{A}}x[NN_{\mathrm{A}}(x) \Rightarrow \exists_{\mathrm{A}}!y[NN_{\mathrm{A}}(y) \wedge SUC_{\mathrm{A}}(x, y)]]$), providing that the subscript 'A' indicates that the relevant statements have to be understood in one of the two previous ways, by taking the relevant version of arithmetic to be A. Although, so understood, these facts wouldn't be related to any particular theory, it remains that the conception wouldn't be faithful to our customary or natural way of understanding what the non-reducible facts about the

numbers 5, 7, and 12, and the natural numbers in general are, since according to that conception these facts are not universal facts about particular versions of arithmetic, but facts about the numbers 5, 7, and 12, and the natural numbers as such.

Moreover, if the facts that 5 + 7 = 12 and that every natural number has a single successor were so conceived, and it were also admitted (in agreement with minimal platonism) that the content of a mathematical theory-tied belief that *p* is that the fact that *p* obtains, it would follow that no proof, within any version of arithmetic, and no argument supporting an axiom of any such theory could, respectively, justify the beliefs that 5 + 7 = 12 and that every natural number has a single successor. Only proofs or arguments within some appropriate theory of arithmetical theories could do this. Hence, since no such theory is available (unless we consider that this is provided for by the historiography of mathematics), it would follow that no justification for these beliefs is available (unless such a justification depends on historiographic remarks). This, once again, is openly unfaithful to our customary and natural conception of what counts as a justification for these beliefs.[24]

It seems, then, that the difficulties that both the platonist and the combinatorialist must face when addressing Benacerraf's dilemma are deeper than worries about whether there are any mathematical objects independently of our intellectual activity, or whether mathematical facts are reducible to proof-theoretical facts. The deeper issue is whether there is room for conceiving mathematical objects as being, as such, independent of mathematical theories, though maintaining that these theories are about them. In other terms: is there room for conceiving these objects as the objects about which these theories are, rather than, merely, as the objects that these theories are about?

There is no doubt that mathematical theories are human constructions. No platonist or realist seems to be able to deny this. At most, one can argue that these constructions are (supposedly) about a transcendent reality. So, wondering whether a mathematical theory M is about objects that are independent of it is tantamount to wondering whether the cognitive subjects that have set up M, or who work in M, or who come to learn M can be credited, while doing this, with a *de re* epistemic access to the objects M is about, rather than merely with a *de dicto* epistemic access to them. In other words, can it be maintained that it is these very objects that the cognitive subjects are dealing with, or can it only be said that these subjects are dealing with these objects while doing mathematics? In still other words, the question is whether mathematical objects can be fixed as individuals we have epistemic access to—i.e. individuals we can distinguish from other individuals, and, when focusing on some of them one by one, also from each other—independently of M, to the effect that we may take them as the objects M is about, and not merely take M to be about them.

It follows that a way of meeting both challenges at once (and indeed the only possible way to do this) is to provide a plausible account of mathematics according to which mathematicians are credited with a *de re* epistemic access to mathematical objects when they devise a mathematical theory about such objects, work within the theory, or merely learn it.

One could think that it is just for granting this that a platonist matching with Field's description maintains that there are mathematical objects, that they are mind- and language-independent, and that mathematics is about them. But this grants, at most, that the same mathematical vocabulary may be used to speak about the very same objects in whatever context of use. It doesn't yet grant that one can have a *de re* epistemic access to them. Since the possibility of such an epistemic access to mind- and language-independent abstract objects is precisely what is called into question by Benacerraf's dilemma. So, if we concede that a platonist matching Field's description is not able to meet this challenge, we must conclude that the very same platonist is not able to account for the possibility of having a *de re* epistemic access to mathematical objects either.

It seems, then, that a platonist can hope to account for this only provided that he or she is not only a minimalist platonist, but is also ready to deny either that mathematical objects are mind- and language-independent—i.e. to deny (ii)—, or that they are abstract—i.e. to deny (iii). If the latter option is discarded as highly implausible, such a platonist should then maintain that mathematics is about abstract objects that we fashion through our intellectual activity—which amounts to denying (ii)—, and look for a way of accounting for our fashioning of these objects that leaves open the possibility of having a *de re* epistemic access to them.

In my view, what is distinctive of an object (either concrete or abstract) is just this: an object is an individual item—that is, an item which may provide the putative reference of a singular term, or to count as an element of the putative range of a first order quantifier (possibly in a multi-sorted first-order language, or in a multi-sorted first-order fragment of a higher-order language)—that some cognitive subjects can have a *de re* epistemic access to. This means that there is, for these subjects, a way of relating to this item such that one may argue that it is with it that these subjects are dealing, and not merely that these subjects are dealing with it. Existence, intended as a primitive intrinsic condition not to be submitted to any sort of specification, isn't what is at stake here: taking *a* to be an object does not require to admit that *a* exists in such a primitive sense. If one is willing to argue that an object exists—as I am willing to as far as mathematical objects are concerned—one has to specify the particular sense in which it does, a sense that depends on the particular nature of the object or, better, on the particular nature one ascribes to it. Fashioning an abstract object, or a domain of abstract objects, consists, then, in fixing an individual item or a domain of individual items, in such a way as to make it possible for cognitive subjects to have a *de re* epistemic access to it (more on this matter in the next section).

Conversely, nothing that has not yet been so fixed, or that cannot be specified so as to allow cognitive subjects to have a *de re* epistemic access to it, may be considered an object. Such a putative object would be a mere logical reification of a concept, or a bundle of properties, a posit resulting from the nominalization of predicates, or from the association of such predicates to names, along with the stipulation that this is enough for ensuring reference. Hale and Wright seem to imply that something like this happens to places in a structure, as defined in *ante rem* structuralism: according to them, by merely giving an "axiomatic description

[…] characterizing a structure" we cannot "do more than *convey a concept*", namely we cannot "induce awareness of an articulate, archetypical *object*, at once representing the concept in question and embodying an illustration of it" (Hale and Wright 2002: 113). I agree. This does not mean, however, that providing an axiomatic description cannot result in, or be part of, fashioning some abstract objects in my sense. It is so, indeed, when this description makes it possible, or contributes to make it possible, for cognitive subjects to have a *de re* epistemic access to the items so fixed, while dealing with them as objects that a theory, which doesn't involve, as such, this axiomatic description, is about.

I have argued so far that one may meet conjointly the challenges that Benacerraf's dilemma addresses to the platonist and the combinatorialist by providing a plausible account of mathematics according to which mathematicians are credited with a *de re* epistemic access to mathematical objects while they devise a mathematical theory, work in it, or learn it. According to such an account, a theorem, axiom, or any other liminal assumption of such a theory is intended as a *de re* description of these objects, or as a *de re* ascription of properties or relations to them. Hence, the content of a belief expressed by such a theorem, axiom, or liminal assumption is just that these objects satisfy these descriptions or conform to these ascriptions. Proving such a theorem, or endorsing such an axiom or liminal assumption is the same as securing these descriptions or providing grounds for such ascriptions. Moreover, insofar as these descriptions and ascriptions are *de re*, other descriptions or ascriptions may be offered concerning the very same objects depending on other theories which are also about them, and any such theory provides a way of speaking of these objects, i.e. of describing them or of ascribing properties or relations to them.

What I have just said on what I take an object to be, and on what fashioning an abstract object or a domain of abstract objects consists in, should make clear that this way of meeting Benacerraf's challenge is perfectly in line with the view that mathematical objects are objects in a proper sense: they are more than mere logically appropriate reifications of concepts or bundles of properties. It is also perfectly in line with the idea that mathematics is not only a human production, namely the result of our intellectual activity, but also that it is self-determining, that it hardly requires the exercise of a mysterious faculty yielding an access to a transcendent reality. Neither does it rely on the independent existence of such a reality. In short, this way of meeting these challenges is in line with the basic proposals of both a platonist and a combinatorialist.

What remains to be explained is how an abstract object, or a domain of abstract objects, is to be fashioned so as to make it possible for us to have a *de re* epistemic access to it, or, more precisely, how one may achieve such fashioning so as to make it possible for us to have a *de re* epistemic access to it while a mathematical theory about such objects is being devised, worked on, or learned. This is indeed a complex question. Let me try, however, before concluding my paper, to outline the direction along which I think it would be possible to respond to it.

## 4.5 Fashioning Abstract Objects and Having a *de re* Epistemic Access to Them

The first thing to be said is that a *de re* epistemic access to abstract objects that we fashion cannot come together with the very act of fixing them. The former must be subsequent to the latter. Hence, the question is not whether and how we can have *de re* access to abstract objects while fixing them, but rather, whether and how we can fix them in such a way that we may later on benefit from *de re* access to them, i.e. describe them *de re*, or ascribe properties or relations to them *de re*. This is perfectly consonant with the picture of mathematics suggested by platonism as described by Field. According to this view, there are mathematical objects before we can speak of them, describe them, or ascribe properties or relations to them. Still, the sense in which we are said to ascribe properties or relations to mathematical objects, and the way such ascriptions interact with descriptions, are quite different in the two cases.

According to platonism as described by Field, we have no other intellectual capacity relative to mathematical objects than that of recognizing their properties or relations. Any assertion about them is quite literally a description, or at least a purported description. We may attribute properties or relations to them only insofar as we introduce appropriate conceptual tools fitted to a description, so that the ascription is properly a part of a description. It follows that there is no intrinsic difference in nature between the intellectual activity consisting in fixing the mathematical objects and the intellectual activity consisting in asserting something about them: in both cases, what is provided is a (purported) description. Typically, we do the former through appropriate definitions and the latter through appropriate theorems. The definitions can be either explicit or implicit. In the former case, they either come after some axioms or liminal assumptions, and are licensed by them (which is generally the case in formal theories), or come before them and complete them (which is generally the case in informal theories). In the latter case, they precisely consist in some axioms or liminal assumptions. There is, then, no other intellectual activity we can perform with respect to mathematical objects than that of defining them, or of stating some axioms or liminal assumptions about them—which is often the same as defining them—and of asserting some theorems about them (once these theorems have been proved). In any event, all we do is (purportedly) describe them. The only difference is that the description provided by a theorem is secured by the description provided by the relevant definitions, axioms or liminal assumptions: the relevant objects cannot but be as the theorem asserts they are, if they are indeed as the definitions, axioms or liminal assumptions say they are. But whether a statement about them is deemed a definition, an axiom, a liminal assumption or a theorem, pertains to a choice that depends only on our ability to identify the properties and relations of the relevant objects, and/or on matters of expository economy.

According to the picture I'm trying to offer, things are much more complex. It is one thing to fix mathematical objects so as to make it possible for us to have a *de re* epistemic access to them at a later stage, and quite another to deal with them while the access is being secured. The latter can be done in three different ways: we may

select some of them among others; we may offer further specifications of some of them by ascribing new properties or relations to them without modifying their nature; or we may acknowledge that they are so and so, i.e. that they have such and such properties or stand to each other in such and such relations. Typically, we fix mathematical objects through appropriate definitions, either explicit or implicit, in the same connexion as before with axioms or liminal assumptions. This doesn't amount to a description but, rather, to a process of constitution. We make these objects available; under appropriate specifications, we can say that we bring them into existence. We may also resort to definitions, either explicit or implicit, in the same connexion as before with axioms or liminal assumptions, while enjoying a *de re* epistemic access to mathematical objects and dealing with them. This is exactly what we do when we select some of these objects among others, or when we further specify them, or some of them, by ascribing to them new properties or relations in the appropriate way. Still, whereas in the former case we merely offer a description of the relevant objects, in the latter we bring their constitution to completion by making it more fine-grained, so to say: accordingly, the attribution is justified in so far as it is not part of a description, but rather of a constitution. Finally, we typically recognize that the mathematical objects we have *de re* epistemic access to are so and so by proving theorems about them. This consists, once again, in providing a description of these objects, although the role of the description isn't to distinguish them from other objects, say by emphasizing some of their features, but to have a finer grained look at them, by making explicit what is only implicit in their definition, or in the relevant axioms or liminal assumptions. In other terms, what we now achieve is a description secured by the constitution, possibly involving previous selections and proper ascriptions.

An example should make my point clearer.

Suppose we define the cardinal numbers as neo-logicists suggest we should do, i.e. as values of a function from concepts to objects, implicitly defined defined by Hume's principle. It is not mandatory to consider that this definition identifies independently existing objects that are recognized, just because of this, as the cardinal numbers. One can take the definition to be a skillful stipulation that merely fixes some abstract items as something that we shall later on be able to identify as such and talk about. This is done without asserting anything about the properties of these objects and the relations between them, over and above their being values of this function. These objects are then fixed without us asserting anything more about them, in particular without us ascribing relations to them. One can then look at them as such, i.e. have a *de re* epistemic access to them, and define some relations on them or on some of them that we have previously selected.

For example, if Hume's principle is stated as a proper axiom added to an appropriate system of second-order logic whose monadic predicate variables are intended to range over concepts, as neo-logicists suggest, one can define the successor relation on cardinal numbers by appealing to one or another among many well-known equivalent formulas of this system of logic, then consider the cardinal

number of the concept $[x : x \neq x]$ and show that the cardinal numbers that bear the weak ancestral of the successor relation to it form a progression.

It seems clear that, by doing this, one is operating on the cardinal numbers as objects that are fixed in advance (through Hume's principle) and that this is made possible because one has a *de re* epistemic access to them: it is on these objects that the relevant strict-order relation is defined, and this consists in the proper ascription of a relation to them. We then rest on this ascription to select the natural numbers among the cardinal ones. This is done by taking zero to be the cardinal number of the concept $[x : x \neq x]$ and the natural numbers to be those cardinal ones that bear to zero the weak ancestral of the successor relation. Finally, it is of these numbers, so selected among the cardinal ones, that we prove that they form a progression, which entails that each of them has a single successor (merely saying that the cardinal numbers stand in a strict-order relation and form a succession doesn't, therefore, provide a suitable account of what is here at issue). This amounts to a description of these numbers, secured by a previous constitution, and involving a previous proper ascription and selection.

Once this is done, one may define an additive operation on the natural numbers and prove additive theorems, for example that $5 + 7 = 12$. Once again, the operation is defined on these very objects and the theorems are proved about them (merely saying that these objects stand in some additive relation doesn't provide a suitable account of what is here at issue).

It might be argued that when operating in this way, one is working within a particular theory, namely Frege Arithmetic, so that one falls prey to the problems we've already considered above, i.e. to the difficulties linked to the justifications of mathematical beliefs coming from proofs within particular theories. But this would be disregarding a crucial fact, namely that Frege Arithmetic is built in order to deal with objects that are fixed in advance, quite independently of many of its components.

One might perhaps hope to argue that: (i) Hume's principle, intended as a proper axiom added to an appropriate system of second-order logic, is nothing but a particular version of a more fundamental principle, namely a clear-cut but essentially informal stipulation assigning the same cardinal number to any pair of equinumerous concepts; (ii) this more fundamental principle is enough for fixing the cardinal numbers as abstract objects in such a way that we may, later on, have a *de re* epistemic access to them. If this is admitted, one should also concede that Frege Arithmetic is built for dealing with objects fixed in advance, so that stating Hume's principle as a proper axiom added to an appropriate system of second-order logic is tantamount to a description of objects that have been fixed in advance so as to allow us to work with them in a convenient way. It would follow that Frege Arithmetic is a theory of objects we have a *de re* epistemic access to independently of this very theory; in particular that we have a *de re* epistemic access to these objects while we build the theory, work within it and eventually learn it.

Moreover, one could also argue that there is a way of informally defining addition on these objects as a binary operation satisfying a number of conditions specified in relation to them. If this is the case, the definition of addition within

Frege Arithmetic reflects a relation that these objects bear independently of it, and there is room for claiming that a proof of an additive theorem within Frege Arithmetic is a justification of a belief whose content is independent of this theory.

On could object that nothing ensures, *pace* Frege, that arithmetic is a theory of cardinal numbers intended as numbers of concepts, and that an arithmetical belief is therefore a belief about them. There is, indeed, no reason for crediting the identification of natural numbers with numbers of concepts with any sort of pre-eminence over other possible identifications, for example over their identification with ordinals or elements in a progression.

I agree that there is indeed no reason for doing this. Still, what I want to argue for, by offering the foregoing example, is not that Hume's principle (either formally or informally understood) fixes natural numbers as they actually are. It is rather that there is a way of fixing abstract objects counting as natural numbers so as to allow us to have a *de re* epistemic access to them while dealing with them within a given version of arithmetic, which means that we have access to them quite independently of such a version of arithmetic, and possibly of any version of arithmetic, or at least of most components of it. There are certainly other ways of doing this. And there is also room for arguing that the exact content of our belief that 5 + 7 = 12, or that every natural number has a successor, isn't fixed once for all, but can vary from context to context, community to community, or, even, subject to subject.[25]

I have neither the space in this paper, nor enough clear ideas in my mind for arguing in favor of one of these possibilities, or for suggesting a possible alternative. My only purpose, here, was to make clear what is, in my view, the crucial and basic challenge that Benacerraf's dilemma addresses to a plausible philosophy of mathematics, and to suggest that there is a possible way out which is compatible with platonism, or, at least, with a platonist spirit.

## Notes

1. This last challenge is possibly not generalizable to any sort of *a priori* knowledge or beliefs, as opposed to the original version of the dilemma, as recently shown, e.g., in Thurow (2013): Sect. 2.
2. Two clarifications are in order. First of all, although Field writes 'mind-dependent or language-dependent', it seems clear that the 'or' counts here as an 'and'; this is confirmed by many other formulations offered by Field (for example in Field 1989: 27). Secondly, although Field's preference is for labelling this view 'mathematical realism', I prefer to use the term 'platonist' and its cognates, since the former term is, more often than the latter, also used in the literature to refer to other views openly concerned with mathematical truth, or more generally with the truth-value of mathematical statements. I also dislike to use the term 'entity' to denote the kind of thing which, according to a platonist, mathematics is about, since using this term suggests that platonism is quite vague about the logical status of what mathematics is about. This is why,

except in quotes, I shall use the logically much more precise term 'object' instead.

3. This reading is quite openly suggested by Field himself when he writes that "one would have to formulate more clearly the claim that our mathematical beliefs are 'reliable', or 'reflect the mathematical facts'" (Field 1989: 26). It is also suggested, e.g. in Burgess and Rosen (1997: 41–42) and in Linnebo (2006: 548–549).
4. This reading is suggested by Liggins (2006: 137) and seems to be confirmed by two other formulations of what Field takes to be the "key point" of Benacerraf's dilemma, which he offers in Field (2005: 77, 81):

> Our belief in a theory should be undermined if the theory requires that it would be a huge coincidence if what we believed about its subject matter were correct. But mathematical theories, taken at face value, postulate mathematical objects that are mind-independent and bear no causal or spatio-temporal relations to us, or any other kinds of relations to us that would explain why our beliefs about them tend to be correct; it seems hard to give any account of our beliefs about these mathematical objects that doesn't make the correctness of the beliefs a huge coincidence.
>
> The Benacerraf problem […] seems to arise from the thought that we would have had exactly the same mathematical […] beliefs, even if the mathematical […] facts were different; because of this, it can only be a coincidence if our mathematical […] beliefs are right, and this undermines those beliefs.

5. Remark that I am not referring here to mere conjectures having a conditional or dubitative content, since one could argue that these are not expressions of genuine beliefs. I refer to unquestionably genuine beliefs expressed by apodictic statements like '$2^{\aleph_0} \neq \aleph_1$' or 'the real part of any non-trivial zero of the Riemann zeta function is 1/2'.
6. This is not the same as arguing that Field's challenge is to be intended as the request that a platonist explain the reliability of the process that is supposed to secure the relevant beliefs. What I mean is that the challenge appears to be plausible only if it is intended as the request that a platonist explain how it happens that a mathematician has the mathematical theory-tied beliefs that *p* only if *p*.
7. Things would be different if the challenge were taken to concern the reliability of the process that is supposed to secure the relevant beliefs. For one could easily admit that, if there is something special that makes this process reliable, then it is thereby available to mathematicians. Considering mathematicians to be the bearers of the relevant beliefs would then be a way of focusing on this special reliability-maker, rather than on the beliefs themselves. Still, I do not think that the problem with Benacerraf's dilemma, however understood, is that of identifying such a special reliability-maker for mathematics (for example the kind of thing that is often allegedly referred to with the term 'mathematical intuition', which is hardly understandable without further specifications).
8. This has already been remarked in Burgess and Rosen (1997: 42) , but see also Linnebo (2006: 571, note 4). Liggins, on the other hand, insists that the two projects of "explaining how our beliefs come to be justified, and […] [of] explaining how our beliefs come to be reliable" are "distinct" and "quite

separate" (and they would be so, even if "being justified" were conceived as being the same as "being formed by a reliable process", since the former project should then involve an explanation of such a conception of being justified, whereas the latter should not), and he also adds that Field's argument "has nothing to do" with the former project, but rather pertains to the latter (Liggins 2006: 139–140; see also Liggins 2010: 73). It is not clear to me what Liggins means by 'reliability of mathematical beliefs.' Still, his point seems to be that one thing is to wonder whether some beliefs count as justified or not, and another is to wonder what in general makes a belief formation process reliable (or, possibly, what in general makes a belief reliable, if the latter question were considered as different from the former). Though I fully agree on this distinction, I disagree that Field's point concerns the question of establishing what makes the formation process of mathematical beliefs reliable, provided that it would be different from the question of explaining how the relevant beliefs reflect the mathematical facts, this having in turn nothing to do with the way the relevant beliefs are justified. Indeed, it seems clear that: (i) for Field, claiming that our mathematical beliefs are reliable means the same as claiming that they reflect the mathematical facts; (ii) Field's point cannot plausibly be made with respect to any mathematical belief, but must be restricted to theory-tied ones. Under such circumstances, separating the explanation of the reliability of the relevant beliefs from any consideration of their justification seems quite artificial, unless it depends on arguing that mathematical theory-tied beliefs are not necessarily justified. Field has suggested something like this when arguing that "many of our beliefs and inferential rules in mathematics, logic, and methodology" are such that "we must be, in a sense, entitled to them by default", and that "our being default-entitled to them" is not to be regarded as a "mysterious metaphysical phenomenon" since what happens "is, basically, just that we regard it as legitimate to have these beliefs and employ these rules, even in the absence of argument for them, and that we have no other commitments that entail that we should not so regard them" (Field 2005: 81–82). According to Field, one reason for considering that, in the case of these mathematical beliefs —namely mathematical axioms or other sorts of liminal assumptions of mathematical theories—"the need for justification doesn't seem as pressing" is that, in mathematics, there does not seem to be a "genuine conflict between alternative theories" because "it's natural to think that different mathematical theories, if both consistent, are simply about different subjects" (Field op. cit.: 82–83). But, as he also observes, this is no "lessening the need for justification"; it merely entails "that the justification for consistent mathematical theories comes relatively cheap: by the purely logical knowledge that the theory is consistent" (Field op. cit.: 83). The latter option (according to which mathematical axioms or other sorts of liminal assumptions of mathematical theories are justified by consistency proofs, or at least by arguing in their favor) fits in perfectly with what I shall say later (though my point also applies if we admit

more substantial sorts of justification for them). Under the former option (according to which we are entitled by default to mathematical axioms or other sorts of liminal assumptions of mathematical theories so that they have no genuine justification at all), something I shall say in what follows would not apply. But the conclusion I shall come to, regarding what I take to be the crucial challenge addressed by Benacerraf's dilemma to a platonist, could be easily restated for it to apply also under this option. More on this in endnote 14.

9. Field (Field 1989: 233–239; Field 1988: 62–67) has considered the possibility of trivially meeting the challenge by observing (in Linnebo's words; see Linnebo 2006: 557) that "the correlation to be explained has no counterfactual force", since mathematics is necessary and mathematical facts obtain in all possible worlds. He has offered different arguments against this line of response, and other scholars have discussed some of them, or offered other arguments to the same effect. I do not wish to enter into this discussion because it seems to me that there is a quite simple way to block a similar response, even if one takes for granted (something I'm not inclined to do) that mathematics is necessary in an appropriate way, so that mathematical facts obtain in every possible world. The point is that, even if it were admitted that mathematical facts obtain in all possible worlds, this would in no way entail that our mathematical theory-tied beliefs are the same in all possible worlds. It is not hard at all to imagine a possible world in which these beliefs include, e.g., the belief that 5 + 7 = 13, though what happens there, as in any other possible world (under the granted assumption) is that 5 + 7 = 12. So, in this setting, a platonist should still explain how it happens that in our actual word our mathematical theory-tied beliefs include the belief that *p* only if *p*, even if this isn't so in all possible worlds (also under the assumption that a mathematical fact obtains in all possible worlds). One could say that, provided it is granted that the fact that *p* obtains in all possible worlds, requiring an explanation of this is not tantamount to requiring an explanation of a genuine correlation. Still, far from solving the problem, this purely terminological remark would leave the problem untouched.
10. It isn't necessary at this point to specify the nature of the relation that mathematical facts are supposed to bear to mathematical objects. Specifications will rest upon metaphysical views which shouldn't affect the points under discussion. What is crucial here is that mathematical facts are taken to depend in one way or another on how mathematical objects are, and on which relation they bear to each others. Following Field's terminology (see, e.g., Field 1989: 232, quoted above in Sect. 4.2), I use 'about' to indicate this unspecified relation: I say that some facts are about some objects to mean that these facts depend on how mathematical objects are, and on which relations they bear to each others.
11. Sereni's conclusion is equivalent to this one only under some specifications. Sereni argues that, when addressed to a platonist, Benacerraf's dilemma faces, as it were, a meta-dilemma structurally quite similar to itself: for it to be recovered so as to avoid begging the question, and being, in this sense, ill-posed, "it should not rely on notions so robust as to make the corresponding

challenge to the platonist prejudicial'; for it to be recovered so as to avoid being confused with other, already well-known charges to platonist views or with the mere requirement that a philosophical account of mathematics meet some basic satisfaction conditions, and so being utterly unspecific, "it should not be so general that no novel or dedicated threat is raised for mathematical platonism." His point is, then, that though "both requirements are desirable and can be defended on their own[…] [,] it is unclear whether they can be satisfied together." Sereni's suggestion is, clearly, that they cannot. As I shall try to explain in what follows, my point is that the former requirement can be fully satisfied so that a genuine challenge may be addressed not only to mathematical platonism, but also to any plausible philosophy of mathematics. That the same challenge could also be detected in other arguments in the philosophy of mathematics is another question altogether: this is certainly true, and far from undermining the problem, it testifies to its deepness.

12. See endnote 11 above for the sense in which I deem the challenge specific.
13. In what follows, when speaking of justifications for mathematical theory-tied beliefs, I shall not be referring to possible arguments, belonging to some abstract domain of arguments that one could take as a justification of these beliefs, but to consensual epistemic practices, i.e. to actual justifications occurring within the relevant theory (or theories), or in relation to it (or them).
14. Of course, if we were merely entitled by default to mathematical axioms or other sorts of liminal assumptions of mathematical theories, so that they had no justification at all (see endnote 8), the challenge would not apply to our beliefs pertaining to these axioms or liminal assumptions. It is, however, easy to restate the challenge so as to make it apply also in this case: what should, then, be explained would be how we could be entitled by default to the belief that a mathematical fact obtains, rather than to the belief that it is appropriate, or perhaps only legitimate, to admit a certain axiom or liminal assumption. After all, under this construal, as Field himself notes, the relevant entitlement by default reduces to the mere circumstance that "we regard it as legitimate to have these beliefs […] [and to have] no other commitments that entail that we should not so regard them" (see endnote 8, again).
15. Burgess and Rosen consider indeed that the last formulation quoted above "is in fact Benacerraf's (writing with Putnam in Benacerraf and Putnam 1983: Introduction, Sect. 9)". What they allude to is possibly the following passage of Benacerraf and Putnam's introduction, where a similar question is raised:

> But why should the simplest and more conservative system (or rather, the system that best balances simplicity and conservatism, by our lights) [that is, the theory we prefer and adopt] have any tendency to be *true*? […] It is hard enough to believe that the natural word is so nicely arranged that what is simplest, etc. by *our* lights is always the same as what is *true* (or, at least, *generally* the same as what is true); why should one believe that the universe of sets […] is so nicely arranged that there is a preestablished harmony between *our* feelings of simplicity, etc., and *truth*?
>
> Benacerraf and Putnam (1983: 35)

This is even closer to an alternative way of posing the challenge that Burgess and Rosen also suggest:

> […] there is a *connection* which has not been explained. It is the connection between set theory's being something that creatures with intellectual capacities and histories like ours might, given favourable conditions for the exercise of their capacities, come to believe, and set theory's being something that is true.
>
> Burgess and Rosen (1997: 47).

16. I'm not sure whether Holland means something like this when he retorts to Burgess and Rosen's understanding of the challenge that this "is not a demand for an external justification of science; rather, it is a demand for the justification of the scientific character of belief formation about abstract mathematical entities" (Holland 1999: 239; see also Linnebo 2006: 552).
17. Notice that this question cannot be appropriately answered by what Linnebo calls 'internal explanation', i.e. by an explanation according to which "mathematicians' tendency to accept as axioms only true sentences is adequately explained by pointing out that the historical process that led to the acceptance of these axioms is a justifiable one according to the standards of justification implicit in the mathematical and scientific community" (Linnebo 2006: 561). Possibly, Linnebo is right in claiming that this explanation is "undefeated", if it is intended to respond to Field's challenge (see Linnebo op. cit.: 563). But it simply does not address my version of the challenge, since what I'm requiring in way of an explanation is just how it can happen that the justifications issued by this historical process (that is, the arguments selected through it, in support of mathematical axioms and other liminal assumptions of mathematical theories, and the proofs within these theories) turn out to be the justifications that some appropriate mathematical facts obtain. An internal explanation in Linnebo's sense no more answers Field's version of the challenge, at least insofar as this is understood as a demand for an explanation of how it happened that this historical process lead mathematicians to (justifiably) have a mathematical (theory-tied) belief that $p$ in case the fact that $p$ obtains. Things go possibly differently with Linnebo's "external explanation", i.e. with an explanation of "*what makes it the case* that the process is reliable", i.e. of "why […] [mathematicians'] methods are conducive to finding out whether […] [mathematicians'] claims are true" (Linnebo loc. cit.). According to Linnebo, in the case of perceptual knowledge, such an explanation should explain the correlation between claims "about physical objects outside of people's sensory surfaces" and methods used "for deciding whether to accept such claims", relying on "the verdicts of […] [people's] senses" (Linnebo op. cit.: 564). This might suggest that what Linnebo is requiring here is just an explanation of the correlation between mathematicians' justifications and mathematicians' claims or, possibly, their (theory-tied) beliefs (Linnebo op. cit.: 569). If this is so, Linnebo comes here very close to my version of the challenge, though he suggests, then, a way to meet it that is quite different from what I shall suggest later on.

18. We can make this clear by pointing out the difference between the setting which underlies this challenge and the setting which underlies a counter-example *à la* Gettier to the tripartite conception of knowledge. One could argue that no such counter-example is available for mathematical knowledge but, if one were possible, it should go as follows. (An alleged counter-example for the case of logical knowledge has been suggested in Besson 2009: 2–4). Suppose that someone—Archy—reads the following sentence in a textbook on number theory: "a prime number is a natural number having no divisor other than 1 and itself." Suppose also that the textbook doesn't offer a definition of natural numbers, taking for granted that these are the well-known numbers 0, 1, 2, etc., and that the definition occurs at page 2, while at page 3, a perfectly usual and universally acceptable definition of the order relation SMALLER OR EQUAL TO on natural numbers is offered. Suppose furthermore that Archy, after having read this (and before coming upon the definition of the strict-order relation SMALLER THAN offered on page 4 and, then, before understanding the difference between an order and a strict-order relation), draws from what he has learned until then, through a simple and perfectly correct deduction, that 1 is a prime number and, then, that there is a prime number smaller or equal to 2. Although there is indeed such a prime number, one shouldn't admit that Archy knows this. One could argue that this is not a suitable counter-example *à la* Gettier by observing that the source of Archy's justification, namely the definition of the textbook, is not an admissible source of mathematical justification. Whether or not this criticism is legitimate, what is relevant here is not whether the counter-example is well-taken, but rather that the setting which underlies it is different from that which underlies the challenge that, in my view, Benacerraf's dilemma addresses to a platonist. In the former, it may not be doubted that the relevant justification is indeed a justification of what the relevant belief is taken to be a belief of. This is simply taken for granted. What is questioned is whether the relevant justification is suitable for turning the relevant true belief into a genuine piece of knowledge. In the latter, things just go the other way around. There is no question whether the relevant justification is suitable for turning the relevant true belief into a genuine piece of knowledge. As a matter of fact, the question doesn't even arise since one doesn't appeal to the notions of truth and knowledge. What is questioned here is whether this justification is a justification of what a platonist takes the relevant belief to be a belief of.
19. Note that such a belief is different from the belief that '*p*' follows from the axioms of the relevant theory, or that it is a theorem of this theory: while this latter belief is a mathematical one (and is even, according to the granted reduction, a prototypical mathematical theory-tied belief), the former is not. The latter depends indeed on the intra-theoretical notion of following from or of being a theorem, while the former is independent of any such intra-theoretical notions and is justified by an empirical examination of a system of appropriate inscription-tokens (together with the admission that such an examination

suffices to justify a belief about the corresponding inscriptions-types). Moreover, the latter contributes to the justification of the former (the converse doesn't hold), and this is precisely the reason why the two beliefs are connected and why a combinatorialist cannot avoid accounting for the way in which the analyses that reveal their respective contents are related.

20. I say 'actual facts' to make it clear that thesis (i) is not compatible with the view that the facts mathematics is concerned with never obtain, since the objects that these facts are about do not exist (a view suggested by Field's arguments in Field 1980 and Field 1982).
21. Note that the negation of (ii), namely the thesis that mathematical objects (if any) are mind- and/or language-dependent, entails (iii), as well as, of course (the subjacent logic, here, being classic), the negation of (iii), namely the thesis that mathematical objects (if any) are concrete, entails (ii), since the idea of concrete mind- and/or language-dependent objects appears inconceivable. Conversely (ii) and the negation of (iii) are, of course, perfectly compatible.
22. See endnote 20.
23. See endnote 21.
24. *Ante rem* structuralism could be taken as the view that mathematical facts are facts about structures, which are, as such, independent of specific theories, or are at least brought about by different theories. For example, according to *ante rem* structuralism, the facts that 5 + 7 = 12 and that every natural number has a single successor could be taken to be facts concerning the structure of a progression, this structure being brought about by any appropriate version of arithmetic. Under this reading, these facts would not be universal facts about versions of arithmetic, but singular facts about a particular structure. A problem with this view is that it either (i) requires that only categorical theories, which all have the same model (under isomorphism), are appropriate rendering of a certain branch of mathematics (for example that only PA2, or other categorical theories having the same model as PA2 are appropriate versions of arithmetic), which is quite implausible, or (ii) depends on a notion of a structure (and on an identity condition for structures) allowing one to admit that different theories, having different models (under isomorphism)—for example PA2, $ACA_0$, $RCA_0$, and FA (provided we only consider the case of arithmetic)—bring about the same structure, which is not what *ante rem* structuralism in Shapiro's and Resnik's versions admits. If, despite this and the difficulty it raises, this second route were taken, it would also be necessary to explain how proofs within a particular theory, among those that bring about the same structure, or arguments related to it, can justify that facts about this very structure obtain, or to provide a general theory of these theories (which certainly couldn't be a general theory of structures), in which justifications for obtaining these facts can be offered. The difficulties in solving these problems set aside (as well as other well-known ones connected to *ante rem* structuralism), it also remains that any possible plausible solution would presumably be, once more, quite unfaithful to our customary and natural conceptions about what counts as a justification of mathematical theory-tied beliefs.

25. One could even imagine that, in some special contexts, the content of these beliefs be entirely theory-laden, i.e. that what one is believing when believing that 5 + 7 = 12 and that every natural number has a single successor be just that some proof-theoretic facts obtain. To my mind, the point, then, isn't to discover or reveal what the real content of these beliefs is, but simply to make it clear that there is room for accounting for the possibility that this content be independent of a particular theory, while maintaining that these beliefs are justified by means of the usual arguments advanced in mathematical practice, which do depend on particular theories. One should also avoid to mistake, e.g., the belief that 5 + 7 = 12 when entertained by a professional mathematician or a mindful user of mathematics—which is indeed about the natural numbers 5, 7, and 12—for the widespread belief that 5 + 7 = 12 when entertained by mathematically uneducated subjects possibly concerned with other objects, such as the numerals '5', '7' and '12' involved in our usual decimal numeral system, or with the countable collections to which these numerals are assigned. It seems clear to me that accounting for this latter belief and its justification is not a task for the philosophy of mathematics proper, but rather a task for some branch of sociology or socio-linguistics. Whatever a philosopher of mathematics might argue for, concerning the belief that 5 + 7 = 12, should indeed be also applicable, *mutatis mutandis*, to other mathematical (or at least arithmetical) beliefs that have no correlate among the widespread beliefs entertained by mathematically uneducated subjects. After all, what philosophy of mathematics is concerned with is mathematics, not the various ways in which mathematical vocabulary is more or less consciously used in everyday life.

## References

Benacerraf, P., & Putnam, H. (1964). (eds.), *Philosophy of mathematics-Selected Readings*, Prentice-Hall, Englewood Cliffs (N.J.), 1964.

Benacerraf, P. (1973). 'Mathematical Truth', *The Journal of Philosophy , 70*, 661–679; also in Benacerraf, Putnam (1964), 403–420.

Benacerraf, P., & Putnam, H. (1983). (eds.) *Philosophy of mathematics-Selected Readings*, 2nd edition, Cambridge University Press, Cambridge, 1983.

Besson, C. (2009). Logical knowledge and Gettier cases. *The Philosophical Quarterly, 59*(234), 1–19.

Burgess, J. P., & Rosen, G. (1997). *A subject with no object: Strategies for nominalistic interpretation of mathematics*. Oxford, New York: Oxford UP.

Field, H. (1980). *Science without numbers. A defence of nominalism*. Princeton, New Jersey: Princeton UP.

Field, H. (1988). Realism, mathematics and modality. *Philosophical Topics, 16*(1), 57–107. Also in Field (1989), pp. 227–281.

Field, H. (1989). *Realism, mathematics, and modality*. New York: Basil Blackwell.

Field, H. (2005). Recent debates about the a priori. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology*, *1*, pp. 69–88. Oxford, New York: Clarendon Press.

Field, H. (1982). Realism and anti-realism about Mathematics, *Philosophical Topics, 13*(1), 45–69.
Field, H. (1989a). 'Realism, Mathematics and Modality', in Field (1989), pp. 227–281. This paper is a substantial reworking of Field (1988).
Hale, B. (1998). Review of Burgess and Rosen 1997. *The British Journal for the Philosophy of Science, 49*(1), 161–167.
Hale, B., & Wright, C. (2002). Benacerraf's Dilemma revisited. *European Journal of Philosophy, 10*(1), 101–129.
Hart, W. D. (1977). Review of M. Steiner, *Mathematical Knowledge*. *The Journal of Philosophy, 74*(2), 118–129.
Heck, R. G. (2000). Syntactic reductionism. *Philosophia Mathematica, 3rd series, 8*(2), 124–149.
Holland, R. A. (1999). Review of Burgess & Rosen 1997 and Shapiro 1997. *Metaphilosophy, 30* (3), 237–245.
Liggins, D. (2006). Is there a good epistemological argument against platonism? *Analysis, 66*(2), 135–141.
Liggins, D. (2010). Epistemological objections to platonism. *Philosophy Compass, 5*(1), 67–77.
Linnebo, Ø. (2006). Epistemological challenges to mathematical platonism. *Philosophical Studies, 129*(3), 545–574.
Linsky, B., & Zalta, E. N. (1995). Naturalized platonism versus platonized naturalism. *The Journal of Philosophy, 92*(10), 525–555.
Linsky, B., & Zalta, E. N. (2006). What is neologicism? *The Bulletin of Symbolic Logic, 12*(1), 60–99.
Resnik, M. (1997). *Mathematics as a science of patterns*. Oxford: Clarendon Press.
Shapiro, S. (1997). *Philosophy of mathematics: Structure and ontology*. Oxford, New York: Oxford UP.
Thurow, J. C. (2013). The defeater version of Benacerraf's problem for a priori knowledge. *Synthese, 190*(1), 1587–1603.
Zalta, E. N. (1999). Natural numbers and natural cardinals as abstract objects: A partial reconstruction of Frege's *Grundgesetze* in object theory. *Journal of Philosophical Logic, 28*(6), 619–660.
Zalta, E. N. (2000). Neo-logicism? An ontological reduction of mathematics to metaphysics. *Erkenntnis, 53*(1–2), 219–265.

# Chapter 5
# A Dilemma for Benacerraf's Dilemma?

**Andrea Sereni**

## 5.1 Fact of the Matter or Matter of Temperament?

A major part of the debate in the philosophy of mathematics of the last forty years has been dominated by attempts at escaping the dilemma Paul Benacerraf suggested in "Mathematical Truth" (Benacerraf 1973). Most attempts have come from mathematical platonists of different varieties, since the dilemma has been perceived to raise a particularly serious epistemological challenge to platonism. Even when some of the assumptions on which Benacerraf relied are discarded, revised versions of the original challenge are thoroughly discussed.[1] Many have offered head-on responses to the dilemma from various directions, but some have questioned in more recent times that any significant, or indeed coherent, epistemological challenge may be recovered from Benacerraf's original dilemma or by its heirs.[2]

My aim here is to offer stronger evidence in support of this last strand of reactions by arguing that there is something *in the very general structure* of Benacerraf-style dilemmas and corresponding challenges to platonism that makes them in some important sense *self-defeating*.

Let me anticipate the basic idea and outline the discussion that follows. One strategy for reacting to the challenge raised by Benacerraf-style dilemmas is to offer a direct response from a platonist standpoint. Under this category fall all those views that Bob Hale and Crispin Wright classify as "conservative," i.e. those which maintain "that pure mathematics is correctly construed at syntactic face value and that, so construed, it represents, at least for the greater part, a body of a priori knowledge" (Hale and Wright 2002: 103). Another strategy is to answer the

A. Sereni (✉)
Scuola Universitaria Superiore IUSS Pavia, Pavia, Italy
e-mail: andrea.sereni@iusspavia.it



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_5

challenge by dropping the ontological assumptions on which it is based. All nominalistic interpretations of mathematics fall in this category.[3] A third strategy is to observe with Burgess and Rosen (1997: 35–60) that most of the arguments available to the anti-platonist in support of the challenge are going to take the form of burden-of-proof arguments and lead to a stalemate between opponents.

Following this last strategy, one might eventually wonder—as suggested in Wagner (1996: 74) whether there is any fact of the matter underlying the question of the existence of abstract mathematical objects, as Peter Strawson had already suggested about a parallel dispute on universals—or whether there is any "vantage point, […] either in the natural world […] or out of it" from where the dispute can be settled, so that any decision will be a matter of "temperament" or "persuasions" (Strawson 1979: 10).

While steering clear of Strawson's temperamental conclusion, I will undertake a similar strategy by suggesting that it is in the very structure of this family of challenges to either (i) be *prejudicial* against their opponent, i.e. to preclude the possibility of any reply—so that, in Strawson's words, "any location of our judgement seat would be a prejudgement of the issue" (Strawson 1979)—, or else (ii) to fail to constitute any substantial threat, i.e. any *novel* challenge with respect to previously known and obvious ones. An excursus through epistemology will provide evidence that the structure of Benacerraf-style arguments and challenges prevents finding a middle position, where the challenge is both substantial and not prejudicial.

In order to show this, I will suggest that the epistemic notions in Benacerraf's original challenge (Sect. 5.2) and the weaker notions later employed (Sect. 5.3)—the former focusing on causal theories of knowledge, the latter on reliabilist theories of justification—both make the argument prejudicial almost to the point of being question-begging against the platonist. I will then suggest (Sect. 5.4) that employing weaker notions merely makes the challenge unchallenging to the platonist (Sect. 5.4). Finally, I will advance a more explicit formulation of the envisaged shortcoming of Benacerraf-style dilemmas, in the form of a (meta-) dilemma that these dilemmas have to confront (Sect. 5.5).

## 5.2 The Original Dilemma and Challenge

### *5.2.1 Knowing the Beasts*

For ease of exposition, let me express a Benacerraf-style dilemma in an enthymematic form, as consisting of two claims: a conjunctive clause, stating that two independent requests should be satisfied together, and a dilemmatic clause, stating

that the satisfaction of either request entails that the other is unsatisfiable. It is the first conjunctive clause that will be of interest here.

Benacerraf's dilemma can be expressed in different ways and with different degrees of generality. One very general formulation is suggested by Benacerraf when he claims that:

> [. . .] the concept of mathematical truth, as explicated, must fit into an over-all account of knowledge in a way that makes it intelligible how we have the mathematical knowledge that we have. An acceptable semantics for mathematics must fit into an acceptable epistemology.
>
> Benacerraf (1973: 667)

However, the "as explicated" clause suggests that a more precise version of the dilemma may be offered, where assumptions on both the semantic and the epistemological horns are made explicit. Since generality of formulation will play some role in what follows, let me start with a more stringent formulation:

[BD$_1$]

(a) A referential (i.e. Tarskian) account of mathematical truth must be compatible with a causal account of mathematical knowledge.
(b) However, the former excludes the latter, and conversely.

I will keep the semantic horn fixed and consider only variations on the epistemological horn. Thus, from now on, by "mathematical knowledge" I will mean "knowledge of mathematical statements under a platonistic (i.e. referential) semantic interpretation."

Clearly, "platonistic" and "referential" are not by themselves synonyms. Mathematical terms could refer to empirical objects. But I will be considering challenges to what Benacerraf calls "the standard view," i.e. the (maybe merely hypothetical) version of platonism according to which mathematical objects are acausal, aspatial, atemporal, and mind-independent. So I will not consider views such as Penelope Maddy's early physicalist platonism.[4]

Causal constraints are distinctive of our knowledge of empirical objects (and of truths about them), and have traditionally been central to naturalistic views in different areas of philosophy. Thus the challenge to the platonist that may be extracted from BD$_1$ is how, if mathematical objects are "the kinds of entities they are normally taken to be" (Benacerraf 1973: 673), we could ever have "natural knowledge" of them, as Hart puts it (Hart, 1977: 126). In Benacerraf and Putnam's more vivid terms, "[we] need some account of how we can have *knowledge* of these beasties, some account of our cognitive relationship to them" (Benacerraf and Putnam [1964] 1983: 31).

Assuming for the time being an intuitive understanding of what a causal theory of knowledge (henceforth CTK) is, the challenge to the platonist can thus be expressed as follows:

[BC$_1$]

(i) A referential semantics entails the existence of mathematical objects as referents of numerical terms (and values of variables) in true mathematical statements.
(ii) CTK is (unrestrictedly) true.
(iii) CTK requires a causal knowledge of the mathematical objects and facts that make mathematical statements true.[5]
(iv) Mathematical objects and facts cannot stand in causal connections with anything, and thus, a fortiori, with human subjects.
(v) Therefore, a referential semantics for mathematical discourse makes knowledge of mathematical truths impossible.

There are at least three hidden assumptions here:

[*Abstractness*] Mathematical objects are abstract (i.e. acausal, non-spatial, atemporal, mind-independent)
[*Knowledge*] We must account for mathematical knowledge.[6]
[*Uniformity*$_{\text{CTK}}$] CTK is the correct theory of knowledge for all areas of inquiry

Benacerraf makes a parallel uniformity requirement for the dilemma's semantic horn, asking for "an overall theory of truth," valid across all areas of discourse, or, equivalently, asking that "the semantical apparatus of mathematics be seen as part and parcel of that of the natural language in which it is done" (Benacerraf 1973: 666). This has been questioned in various ways, but it is not relevant for present purposes: we are only concerned here with the uniformity requirement for the selected theory of knowledge.[7]

### 5.2.2 Causal Knowledge and Abstract Objects

A proper assessment of the strength of BD$_1$ and BC$_1$ requires that the details of the causal theory of knowledge be made explicit. For a working definition of CTK, we may take either Benacerraf's own:

[CTKB]

I favor a causal account of knowledge on which for *X* to know that *S* is true requires some causal relation to obtain between *X* and the referents of the names, predicates, and quantifiers of *S*.

Benacerraf (1973: 671)

or Alvin Goldman's well-known definition, to which Benacerraf himself refers:

[CTKG]

*S* knows that *p* if and only if the fact that *p* is causally connected in an "appropriate" way with *S*'s believing *p*.

Goldman (1967: 369)

Appropriate causal connections are sustained by processes including perception, memory, causal chains in which inferences participate but which originate either in perception or in memory, and combinations of these. Notice that CTK is a local theory: it establishes whether *S*'s belief that *p* counts as knowledge by reference to the particular causal connection between *S* and the fact that *p*. We will see that this marks a difference with (some) reliabilist theories of justification.

#### 5.2.2.1 Incompatibility of CTK and Platonism

Many are likely to agree with Øystein Linnebo that "it is natural to protest that Benacerraf's considerations are biased against mathematics at the very outset" (Linnebo 2006: 546). But one might want to argue for more, namely that CTK makes platonism a non-starter, that it immediately rules out platonism as an incoherent position which posits unknowable objects and unknowable truths: $BC_1$ may then be seen as prejudicial against the platonist. In order to show this, one needs to argue for the following claim:

[*Incompatibility*$_C$] CTK entails that platonism is false.

This is the natural reading of the following passage by Benacerraf:

If, for example, numbers are the kind of entities they are normally taken to be, then the connection between the truth conditions for the statements of number theory and any relevant events connected with the people who are supposed to have mathematical knowledge *cannot be made out* [emphasis mine].

Benacerraf (1973: 673)

Once this is motivated and accepted, one may eventually argue that since [*Incompatibility*$_C$] is, if not immediately evident, at least almost so, the following holds:

[*Prejudiciality*$_C$] Assuming CTK in $BC_1$ makes the latter prejudicial (if not question-begging) against the platonist.

Notice that [*Incompatibility*$_C$] rests on the following claim:

[*Causality*] Causal constraints on knowledge or justification rule out direct knowledge of acausal objects, knowledge of truths about acausal objects, and justification of truths about acausal objects.

Now, in order to argue for *Prejudiciality*$_C$ one also needs to argue for:

[*Hasty Generalization*] Even though causal constraints in accounts of knowledge or justification are adequate for empirical knowledge, the extension of the corresponding accounts to the domain of non-empirical knowledge is the outcome of an unjustifiable generalization.

I will assume *Hasty Generalization* to be true. There seems to be in principle no reason for extending a model of knowledge that has proved appropriate in one area of inquiry (e.g. sensory experience), however crucial to our entire body of knowledge, to other areas whose characteristic facts are presumably radically different (unless that model is so general and schematic as to allow for significantly diverse specifications in different domains).

Let me now briefly offer evidence in favor of [*Incompatibility*$_C$], [*Prejudiciality*$_C$] and the part of [*Causality*] concerned with knowledge. (I shall consider the part of [*Causality*] concerned with justification in the following section.)

#### 5.2.2.2 Reactions to CTK

Different reactions to the appeal to CTK in BC$_1$ are possible on the platonist side. The platonist might attack premise (ii) and claim that:

(a) CTK is *not* valid across all areas of inquiry. At the very least, we lack sufficient reasons for believing that it is.

Burgess and Rosen have convincingly argued that any attempt to claim the contrary is bound to end up in a stalemate (Burgess and Rosen 1997: 35–60).[8] Be that as it may not be, [*Hasty generalization*] might be used in justifying (a), although it may not sufficient.[9] In any event, accepting (a) stands clearly in no contrast with either [*Incompatibility*$_C$] or [*Prejudiciality*$_C$]. Premise (ii)—in case no other premise is questioned—precludes mathematical knowledge.

Alternatively, the platonist might also want to claim that:

(b) CTK does not necessarily require a causal knowledge of mathematical objects and facts that make mathematical statements true.

Following a similar line of thought, Hale (1987: Chap. 4) suggests that two different forms of CTK should be distinguished. There is a strong CTK (CTK$_S$) which is incompatible with platonism and is by itself untenable. CTK$_S$ allows Hale to reject premise (iii). There is a weak CTK (CTK$_W$) which is compatible with platonism insofar as it does not require a direct causal connection between mathematical facts and corresponding beliefs. Hale suggests that CTK$_W$ is better treated as a theory of justification than as a theory of knowledge, and I will discuss it as such below. On the other hand, the fact that CTK$_S$ is incompatible with platonism is clearly no obstacle to either [*Incompatibility*$_C$] or [*Prejudiciality*$_C$].[10]

Alternatively, the platonist might want to reject (iv) by claiming that:

(c) it is possible, or at least conceivable, that mathematical objects and facts stand in causal connection with human subjects.

Whether possibility or conceivability is at stake will depend on the reading of "can" in premise (iv). Benjamin Callard has suggested that Benacerraf should be interpreted as claiming that the required causal connections are unconceivable. He then argues that, even on the assumption of CTK, it is not "unintelligible how we can have mathematical knowledge whatsoever" (Callard 2007, quoting from Benacerraf 1973: 662). This would tell against (iii), [*Incompatibility*$_C$] and [*Prejudiciality*$_C$] and would be the beginning of an answer to BC$_1$. So let us briefly consider his arguments.

#### 5.2.2.3 Is Platonism Conceivable?

Callard argues that there is nothing unintelligible in there being efficient causation from abstract objects to our brain and mental states. Two features of causation seem to rule this out: the requirement of contiguity relations between causes and effects, and the idea of reciprocal action between objects in the form of energy transfer. Abstract objects are allegedly both non-located and unchanging, and this seems respectively incompatible with both of these features. But Callard argues that:

> Efficient causal relations unsupported by contiguity relations are perfectly intelligible and (therefore) apparently possible, even if they are not actual,

and that

> […] [t]here is no contradiction, or any other conceptual or metaphysical difficulty, in accepting the claim that abstract objects impart energy to us, and thereby change us, without themselves receiving any energy or suffering any change

to conclude that:

> […] contiguity relations and reciprocal action are contingent features of efficient causation; so their strong impossibility in the case of abstract objects cannot support the claim that efficient causal relations with abstract objects are strongly impossible.
>
> Callard (2007: 350, 351, 352)

Jody Azzouni raised two objections to Callard's proposal. They are based on the following assumption:

> Some explanation of how it is possible for the different effects that different [p]latonic objects have on different objects in space and time is — in principle (a priori, if you will) — required if we are to claim that [p]latonic objects can have causal effects.
>
> Azzouni (2008: 398)

But the difference between, say, 13 causing me to think of it, and 13 causing a perfect duplicate of myself to think of it some minutes later, seems to be localizable

only in spatio-temporal differences between me and my duplicate, since "temporally and spatially, 13 has the same relationship to both of us" (Azzouni 2008). So we need to postulate what Azzouni calls "brute 'abstracta sensitivity relations'" (Azzouni 2008: 399) and this postulation seems unjustified. If this is correct, Callard has not shown that causal relation with abstract objects is intelligible.

I agree with Azzouni's objections, but I think they concede yet too much to the opponent. Let us grant that "contiguity relations and reciprocal action are contingent features of efficient causation." Callard does not say which essential and intrinsic features causation is supposed to have, over and above these contingent ones. Without this explanation, Callard's claim seems to reduce to the claim that there is some sort of relation, call it causation*, that could hold between physical objects, e.g. human brains, and abstract objects such that:

(i) causation* is a relation connecting human subjects and abstract objects in a knowledge-supporting way (i.e. it can occur as a replacement of causation in standard definitions of knowledge),
(ii) causation* is a direct form of acquaintance with abstract objects.

This makes knowledge by causation* dangerously close to knowledge by mathematical intuition. But it is not just because it satisfies (i) and (ii) that intuition can be said to be causal in character. The problem here is not merely that the proper causal character of the notion is now missing, but that it is not clear why, in connection with non-abstract objects, we should ever have come to use a notion like causation*, which is explained in terms of relations with abstract ones.

Suppose one nevertheless finds an answer to this problem. It still isn't clear what notion of causation we are left with. What is a necessary condition for causation? I see only one possible reply.[11] Callard often speaks of objects "affecting us," or "effecting changes on us." He could then reply that the essential feature of causation is that for *a* to stand in a causal relation with *b* is for *a* to effect changes in *b*. But this seems to be inadequate: we still lack an explanation of what "effecting changes" means (on top of Azzouni's remark that we miss an explanation of how different effects might be caused). We may of course know what it means for us to *undergo* changes when we form beliefs, e.g. the belief that 17 is a prime number, by describing the change in our cognitive and, arguably, neuro-biological states. But we still lack any explanation of how 17 could ever *effect* this change in us.[12] Unless this explanation is provided, we have not been shown that it is conceivable, let alone possible, that abstract objects effect changes on us.

So Callard seems to leave us either with what looks like an utterly impoverished notion of causation, or with no explanation of causation at all. Even if "contiguity relations and reciprocal action are contingent features of efficient causation," this is a far cry from claiming that we have been shown that causal relations with inert abstract objects are conceivable.

### 5.2.3 Conclusions to Section 5.2

From the previous section it follows that, if one assumes CTK in $BC_1$ (i.e. premise (ii)), premise (iv) will follow, since by the very definition of causation no abstract object can stand in causal connections with a human subject (and, *a fortiori*, in knowledge-conferring causal connections with a human subject). Premise (iii) seems uncontroversial, though we have left some provisos for later.

We have not been shown, therefore, that CTK is in any clear sense compatible with platonism. Assuming CTK precludes a knowledge of abstract objects, and makes $BC_1$ prejudicial, up to the point of being question-begging, against the platonist.

## 5.3 The Revised Dilemma and Challenge

### 5.3.1 The Challenge Revisited

One may argue, even if one accepts the conclusions of Sect. 5.2, that Benacerraf was pointing to a more general challenge to platonism, one that would stand even when notions weaker than causation are involved, and then offer revised versions of $BC_1$ accordingly. The best-known revised form of $BC_1$ is due to Hartry Field.

Field stresses that "Benacerraf's formulation of the challenge relied on a causal theory of knowledge which almost no one believes anymore," but also that "he was on to a much deeper difficulty for platonism" (Field 1989: 25). He adds that Benacerraf's challenge may be stated in such a way that it "does not depend on any assumption about necessary and sufficient conditions for knowledge," showing that it "depends on the idea that we should view with suspicion any claim to know facts about a certain domain if we believe it impossible in principle to *explain* the *reliability* of our beliefs about that domain [emphasis mine]" (Field 1989: 233). Field's epistemological argument is not meant to prove that platonism is false, but is intended to "raise the costs" of giving a platonist interpretation of mathematics. His revised challenge can be reconstructed as follows:

[$RBC_1$]

(i) [*Initial Plausibility*] We are (initially) justified in believing in the existence of mathematical objects as being those (abstract) objects our mathematical statements are about.
(ii) [*Reliability Claim*] Mathematicians' mathematical beliefs are reliable.
(iii) [*Explanation*] The platonist must explain the [*Reliability Claim*].
(iv) [$Impossibility_R$] It is not possible for the platonist to explain the [*Reliability Claim*].
(v) [*Defeat*] [$Impossibility_R$] defeats [*Initial Plausibility*].

Presumably, the platonist will accept (i). Field grants that plausibility can derive from indispensability considerations (and even from evolutionary considerations).

Platonists will also likely accept (ii), assuming that expert mathematicians tend, most of the time (read: for most *p*s), to have true mathematical beliefs: they will believe that "the mechanisms" underlying the formation of our beliefs about mathematical objects can "so well reflect" the facts about them (see Field 1989: 26). Premise (iii) amounts to the quite acceptable claim that "special 'reliability relations' between the mathematical realm and the belief states of mathematicians" should not be accepted as 'brute facts'" (Field 1989: 26–27).

The challenge, for the platonist, is to offer the explanation required in (iii) and Field's allegation is expressed by [*Impossibility*$_{R}$], which undermines initial plausibility for platonism.[13]

Notice that an alternative version of [*Reliability Claim*] would read the claim that mathematicians's mathematical beliefs are reliable as the claim that if mathematicians accept *p* as an axiom, then *p*. This is so because we have a plausible and easy explanation of the reliability of our beliefs about theorems: they are reliable because theorems are obtained by proofs from accepted and consistent axioms.[14] Thus the thesis which calls for an explanation may be restricted to our beliefs in axioms. Field's RBC$_1$ is meant to be superior and stronger than BD$_1$ despite the alleged similarity in structure,[15] insofar as it "can be put without use of the term of art "knows," and also without talk of truth [. . .]" (Field [1988] 1989: 230). As we have seen above, it is also meant to be more innocent and neutral on matters semantic and epistemological. More or less explicitly, Field suggests the following claims:

(a) RBC$_1$ need not assume any "heavy-duty" notion of truth over and above a disquotational one, and may even be formulated without talk of truth at all,
(b) RBC$_1$ makes no use of the terms "know," "knowledge" and their cognates,
(c) RBC$_1$ doesn't appeal to any particular epistemological theory,
(d) RBC$_1$ does not appeal to any epistemic notion,
(e) RBC$_1$ doesn't concern the justification of mathematical beliefs but only the explanation of their reliability.

There are, however, several reasons for being skeptical about Field's declarations of neutrality. Let us consider these in turn.

Point (a) is the least controversial; it isn't however of present interest. Point (b) seems to mark a genuine difference between BC$_1$ and RBC$_1$. No appeal to any particular theory of knowledge seems required by RBC$_1$. Points (c), (d) and (e) are more controversial and indeed relevant: if they are correct and the platonist is not able to meet the challenge, platonism will have been proven untenable on quite minimal grounds, and the challenge will turn out to be both substantial and not prejudicial. But they can be questioned.

First of all, they can be questioned on sociological grounds. Point (c) is striking indeed: to the extent that causal theories of knowledge are representative of the anti-rationalist, empiricist and naturalist philosophical trends that were pervasive when Benacerraf first proposed his dilemma, reliabilist theories of justification were also representative of the very same trends at the time Field was reformulating it. In both cases, Goldman's work has a major relevance. CTK was first fully expounded in

Goldman (1967), whereas the reliabilist theory of justification (henceforth RTJ) was first presented in Goldman (1979). Any reader familiar with the epistemological debate of those years might have been expected to associate Field's mention of reliability with Goldman's views and thus to take the fact that $\text{RBC}_1$ appeals indeed to a particular epistemological theory as evidence against (c), and to then take as evidence against (d) the fact that, as a consequence, $\text{RBC}_1$ appeals to distinctively epistemic notions, if not explicitly—if reliability itself is not considered an epistemic property—at least surreptitiously, when the use of "reliability" is taken as pointing to reliably formed, i.e. justified beliefs. Sociological arguments, however, would at best point to a possible (voluntary or accidental) underestimation of the epistemological debt of the relevant notions in Field's own exposition.

Field suggests that $\text{RBC}_1$ is only concerned with the explanation of reliability, and not with justification, which he takes to be distinct from the former:

> Claims of initial plausibility are of some help to the platonist […] in answering questions about the justification of particular mathematical beliefs […]. But to give them a justificatory role does nothing to explain the reliability of this class of judgements.
>
> Field (1989: 28)

However, as Burgess and Rosen have noticed, the gist of the challenge seems straightforwardly to be that if reliability is not explained, then our mathematical beliefs will turn out to be unjustified, contrary to initial plausibility considerations:

> Though Field maintains that his challenge is "not to our ability to justify our mathematical beliefs," and though he does not explicitly rely on any causal theory of justification, the implicit suggestion in [*Explanation*] can hardly be anything but this, that if the reliability thesis cannot be explained, then continued belief in claims about mathematicalia is unjustified.[16]
>
> Burgess and Rosen (1997: 42)

This goes against (e).

True, as David Liggins has pointed out in a discussion of related issues, we should distinguish between explaining how our beliefs come to be reliable and explaining how our beliefs come to be justified. For even once we have the former, we need something more to have the latter: we need to "add the assertion that being reliably formed suffices for justification" (Liggins 2006: 139n4). One could appeal to this distinction to defend Field on point (e): $\text{RBC}_1$ only requires explaining reliability, not justification.

But if this is so, then it is not clear that $\text{RBC}_1$ encodes any sensible challenge to the platonist—not because it asks for too much, but because, on the contrary, it asks for too little. On one interpretation of reliability according to which reliability is the property of a process of producing true beliefs most of the time, based on actual past occurrences (see Sect. 5.3.1.2), there will be no possibility of distinguishing between the reliability of mathematical methods and the alleged reliability of some unreliable procedure hitting on true mathematical beliefs most of the time (like that adopted by Linnebo's (2006) Lucky Fool, who forms mostly true mathematical beliefs just by tossing a coin).[17] If, on the other hand, explaining reliability involves

explaining what it is for a method or process to make it probable that, even in future and counterfactual instances, it will produce mostly true beliefs, then it is not clear that there is any substantial distinction left between a belief being formed by a reliable process and a belief being justified. The point is rather that, independently of whether one explains reliability in causal or non-causal terms, forming beliefs by a process such that we can tell which features of the process will make these beliefs mostly true in actual and non actual situations seems *ipso facto* equivalent to being justified in our holding these beliefs.[18]

To sum up, if RBC$_1$ merely asks for an explanation of reliability as unrelated to justification, it will be immaterial to the relevant debate: meeting that challenge, easy or hard as it may be, will not provide any epistemic defence of platonism. Point (e) seems preposterous: a minimal condition for RBC$_1$ to pose a significant challenge is that an explanation of reliability contribute to an explanation of justifiedness. This also tells against (d). RBC$_1$ should better be understood as asking not merely for an explanation *that* mathematicians' beliefs are reliable, but rather an explanation *why* they are reliable and thereby justified.

#### 5.3.1.1 Field, Reliability, and Causation

Let us grant, then, that Field is appealing to some epistemically relevant notion of reliability. The question is: how should reliability be explained? If reliability is to be explained in causal terms, it would again make RBC$_1$ prejudicial against the platonist. Is Field assuming that a proper explanation of reliability should be given in causal terms? Here is what Field says (bracketed letters are for ease of reference):

> [A] We need an explanation of how it can have come about that mathematicians' belief states and utterances so well reflect the mathematical facts. But there seems prima facie to be a difficulty in principle in explaining the regularity. [B] The problem arises in part from the fact that mathematical entities, as the platonist conceives them, do not causally interact with mathematicians, or indeed with anything else. This means that we cannot explain the mathematicians' beliefs and utterances on the basis of the mathematical facts being causally involved in the production of those beliefs and utterances; or on the basis of the beliefs and utterances causally producing the mathematical facts; or on the basis of some common cause producing both. [C] Perhaps then some sort of non-causal explanation of the correlation is possible? Perhaps; but it is very hard to see what this supposed non-causal explanation could be. Recall that on the usual platonist picture, mathematical objects are supposed to be mind- and language-independent; they are supposed to bear no spatio-temporal relations to anything, etc. The problem is that the claims that the platonist makes about mathematical objects appear to rule out any reasonable strategy for explaining the systematic correlation in question.
>
> Field [1988] (1989: 230–231)

This is how Field argues for [*Impossibility*$_R$]. Divers and Miller read this passage as the conjunction of the two following claims (see Divers and Miller 1999: 278–279):

(a) Any causal explanation of reliability is incompatible with the acausality of mathematical objects.
(b) Any non-causal explanation of reliability is incompatible with the mind-independence of mathematical objects.

(b) is a questionable reading of Field's words. First of all, it isn't clear that mind-independence alone is incompatible with non-causal explanations of reliability. In [C], the failure of spatio-temporal location is put on a par with mind-independence[19]; *both*, together with *other* features ("etc."), are said to preclude non-causal explanations of reliability. Field's dilemma seems better expressed as follows: (platonistic) mathematical objects are characterized by a bunch of features (acausality, aspatiality, atemporality, mind-independence); among these, (a*) acausality precludes causal explanations of reliability; and (b*) the remaining ones preclude non-causal explanations of reliability.

While there is a clear hint for a possible argument underlying (a*), no such hint is available for (b*): nothing does, not even *prima facie*, rule out *any* explanation of reliability just because the entities involved are atemporal, aspatial, and mind-independent (think of Fregean or neo-Fregean epistemologies, or of a believer's religious beliefs, or perhaps of our beliefs concerning time). (b*) seems merely to assume that no other sort of explanation could be hoped for; this is an unmotivated expression of skepticism (and, to some extent, a variation on [*Hasty generalization*]). Field was surely aware that (neo-logicists) candidates for non-causal explanations of reliability were on the market.[20] His discussion seems then to suggest the underlying thought that *the only plausible explanation* of reliability is causal in character.

On the other hand, claim (a) seems correct: causal explanations of reliability does preclude the possibility of reliable beliefs about acausal objects. But this should be now argued in more details than Field suggested, with the additional help of some recent works on reliabilism.

#### 5.3.1.2 Reliability in the Epistemological Debate

How do reliabilist theories fare with respect to the issue of platonism? Reliabilist constraints on knowledge were initially suggested by Ramsey (1931) and later arose as means of avoiding accidentality (see Unger 1968) or luck in the way a subject $S$ comes to have a true and justified belief that $p$, in response to Gettier cases.

A first version of reliabilism was offered as a theory of (non-inferential) knowledge in Armstrong (1973): the so-called "reliable-indicator" theory (or "thermometer-theory"). According to Armstrong, a subject $S$ has a non-inferential knowledge that $p$ iff $S$ is such that there is a law-like connection in nature such that if $p$ is true, then $S$ believes that $p$.[21] Subjects might under appropriate circumstances be nomologically reliable indicators of the holding of certain facts, just like thermometers are under appropriate circumstances nomologically reliable indicators of external temperature being such-and-such.

We need not enter here into the details of Armstrong's theory, since asking for some law-like connections holding in nature between mathematical facts or objects and human beliefs seems to ban a platonist interpretation of mathematics from the outset. Notice nevertheless that (i) Armstrong suggests a non-platonist interpretation of mathematics; (ii) Armstrong conceives of mathematical beliefs as general beliefs (even when it comes to existential statements) and thinks this allows him to conceive of all mathematical knowledge as *inferential*; (iii) as a consequence of (ii), the definition just rehearsed is not meant by Armstrong to apply to mathematical knowledge; (iv) despite Armstrong's own views about mathematics, we are assuming that inferential knowledge alone cannot suffice for a comprehensive account of mathematical knowledge: we also need non-inferential mathematical knowledge, and the definition above should then apply to it too.

The best developed form of reliabilism is due to Goldman. Goldman (1967) originally introduced CTK by suggesting that the justification condition be dropped from the JTB definition of knowledge and substituted with the condition that the fact that $p$ be causally connected in an "appropriate" way with $S$'s believing that $p$.

Unfortunately, CTK too fell prey to Gettier-style counterexamples, most notably to the fake barns mental experiment.[22] This led Goldman (1976) to Provide a modified theory of knowledge that will be recalled later (cf. sect. 4.1.1 below). The first suggestion for a reliabilist condition on knowledge was suggested in Goldman (1975):

> […] the causal theory of knowing does not say that any causal connection between the fact that $p$ and $S$'s belief yields knowledge; the theory requires that the causal connection be an "appropriate" one. But in order for a particular causal connection to be appropriate, it is sufficient, I think, that it be an instance of a kind of process which *generally* leads to true beliefs of the sort in question. […] [B]oth the chicken-sexer and the rain-predictor *have reliable techniques* for forming beliefs about their respective subject matters, even though neither has any idea what his technique is, and even though neither may *know* that his technique is perfectly reliable.
>
> Goldman (1975: 116)

Goldman made it clear in later works that the reliability constraint does not require that the believer and the relevant fact be causally connected. Knowledge is still explained in causal terms, by means of "causal mechanisms" that are "in an appropriate sense" reliable:

> Like an earlier theory I proposed, the envisaged theory would seek to explicate the concept of knowledge by reference to the causal processes that produce (or sustain) belief. Unlike the earlier theory, however, it would abandon the requirement that a knower's belief that $p$ be causally connected with the fact, or state of affairs, that $p$.
>
> Goldman (1976: 771)

Goldman also expands on which causal processes or mechanisms would count as reliable:

> What kinds of causal processes or mechanisms must be responsible for a belief if that belief is to count as knowledge? They must be mechanisms that are, in an appropriate sense,

> "reliable." Roughly, a cognitive mechanism or process is reliable if it not only produces true beliefs in actual situations, but would produce true beliefs, or at least inhibit false beliefs, in relevant counterfactual situations.
>
> Goldman loc. cit.

Reliabilism is given a proper definition in Goldman (1979) as a way of understanding the notion of justification within the framework of CTK. In Goldman (1979: 1), Goldman revises his earlier (1967) claim that justification should be dropped by means of the JTB definition of knowledge, arguing that what has to be dropped is "Cartesian" justification (the current time-slice capacity of giving accessible-to-the-subject reasons in support of one's own beliefs), whereas under RTJ, justification "is necessary for knowing, and closely related to it."

A standard reliabilist definition of justification is the following:

> If *S*'s believing *p* at *t* results from a reliable cognitive process (or set of processes), then *S*'s belief in *p* at *t* is justified.
>
> Goldman (1979: 13)

Goldman defines his process-reliabilist account as "genetic," since for a belief to be justified is not something which concerns the doxastic state of *S* at *t*, but rather the history of the formation of *S*'s belief that *p*. This account concerns global reliability: it is not centered on the particular connection between the believer and the particular belief that *p*, but rather on the overall (global) reliability of the process by which *p* is formed. Reliability, according to Goldman (Goldman 1979: 11), "consists in the tendency of a process to produce beliefs that are true rather than false." "Tendency" may be understood as either "actual long-run frequency" on past records, or "propensity" on past and possible future occasions.[23] Standard examples of reliable processes are standard perceptual processes, memory, good reasoning, introspection; standard examples of unreliable processes are confused reasoning, wishful thinking, reliance on emotional attachment, mere hunch or guesswork and hasty generalization.

#### 5.3.1.3 Is Goldman's Reliability Explained in Causal Terms?

Can we say that Goldman's explanation of reliability makes no essential appeal to causal relations? The simple answer to this question is: no. There is plenty of evidence that reliability is conceived by Goldman *in causal terms*. For instance, after having reviewed a number of classical definitions of justification—appealing to infallibility, self-evidence, self-presentation, incorrigibility, etc.—, Goldman explains why he found all them wanting as follows:

> Notice that each of the foregoing attempts confers the status of "justified" on a belief without restriction on why the belief is held, i.e., on what *causally initiates* the belief or *causally sustains* it. […] I suggest that the absence of causal requirements accounts for the failure of the foregoing principles.

> […] [C]orrect principles of justified belief must be principles that make causal requirements, where "cause" is construed broadly to include sustainers as well as initiators of belief […].
>
> Goldman (1979: 8–9)

He claims later on:

> The justificational status of a belief is a function of the process or processes that cause it, where (as a first approximation) reliability consists in the tendency of a process to produce beliefs that are true rather than false.
>
> Goldman (1979: 10)

Goldman defines a process as a "*functional operation* or procedure, i.e., something that generates a *mapping* from certain states—'inputs'—into other states —'outputs'" (Goldman 1979: 11). In the present case, outputs are states of believing this or that proposition at a given moment. The question that concerns us here is whether any such process could have inputs provided by abstract mathematical objects or facts about them. Still, both then and now (see e.g. Goldman 2011: 9) Goldman is clear that reliable cognitive processes, even those in which inference is involved, must constitute a "chain" that "must ultimately terminate in reliable processes having only *non-doxastic inputs*, such as perceptual inputs [emphasis mine]."

But how could we have non-doxastic inputs about abstract mathematical objects? One possibility, of course, is that of postulating some special faculty of mathematical intuition, allowing for some sort of quasi-perceptual contact with mathematical objects. I will assume that the prospects for a full-fledged account of such an alleged faculty are lame. In any case, it seems that non-doxastic inputs amenable to enter into causal cognitive processes could only come from sources capable of transmitting informations to us; and again, it is wholly unclear how abstract objects could be supposed to do that.

In any case, cognitive processes in Goldman's sense are supposed to have a psychological plausibility that the alleged faculty of intuition lacks. Clearly, this is not to say that there is no experimental evidence for the existence of dedicated cognitive skills for numerical cognition. Simply, there is no direct evidence that these skills are fit for the detection or description of abstract mathematical objects.

Others have already extensively argued that most common reliabilist conceptions of either knowledge or justification cannot be coupled with the postulation of mathematical intuition. In particular, Albert Casullo has suggested that there is sufficient evidence in favor of the following two claims:

(A) all basic reliable belief forming processes involve the objects of belief as a cause of the belief (this being argued on the basis of inductive evidence from known reliable processes)
(B) there can't be any basic psychological processes that generate beliefs about objects which are causally inert

and that these together imply that no intuitional knowledge of mathematical abstract objects is compatible with reliabilism (Casullo 1992: 582–583).[24]

#### 5.3.1.4 Moderate Naturalism, the A Priori, and Mathematical Knowledge

How does Goldman account for mathematical knowledge? If we look at how Goldman (1986) considers mathematical knowledge, we see that he has in mind either non-platonistic views of mathematics (Hilbert's, Kitcher's), or Maddy's physicalist platonism. More generally, he discusses mathematical cognition (and the role that visual experience can have in it), rather than the content of mathematical statements.

Recently, however, Goldman has argued for a Moderate Naturalism (MN) which is meant to accomodate a priori knowledge (see Goldman 1999).[25] It is a common platonist claim that mathematical truths are a priori. Should we conclude that Goldman's MN, which still vindicates a form of reliabilism about justification, allows for knowledge of mathematical claims platonistically construed?

Among other things, Goldman accepts that it is a mark of the a priori that a priori knowledge, or warrant, has a "non-experiential, i.e., a non-perceptual source or basis" (Goldman 1999: 4–5).[26] If his MN allows for a priori knowledge or warrant, it will also allow for belief-forming processes which have a non-experiential, non-perceptual basis. If it does, then Goldman would be acknowledging the existence of reliable processes with a non-doxastic, non-experiential and non-perceptual basis. Goldman seems to take this liberal aspect of his naturalism as a way of accommodating mathematical knowledge. He claims:

> Another feature of rational insight, however, might frighten off naturalists. This is the perceptual model of rational insight, in which the objects of rational insight are somehow cognized in a fashion analogous to the perception of physical objects. Perception is a causal process, in favorable cases, a process that causally connects a perceived object with the perceiver's mental experience. If rational insight is understood on this model, it must consist in a causal connection between the realm of rationally knowable objects and the knower's cognitive awareness. But it is highly doubtful, from a naturalistic perspective, that such causal connections could obtain. Benacerraf (1973) crystallized this problem in the domain of numbers. If numbers are [p]latonistic entities, can they really have a causal connection with people's mental lives? This problem seems particularly threatening to the form of naturalism adopted here, because (MN) endorses a causal theory of warrant. How can this form of naturalism be reconciled with a priori knowledge?
>
> A crucial step in the reconciliation is to distinguish two types of causal processes, what I shall call *intra-mental* processes and *trans-mental* processes. Intra-mental processes occur wholly within the mind; trans-mental processes include links that are external to the mind as well as links that are internal. Warrant-conferring processes, as envisaged by (MN), are intra-mental processes; they don't encompass objects outside the mind (although the contents of their constituent states may refer to such objects). Thus, a priori warrant does not require the sort of trans-mental, perception-like process that Benacerraf was discussing.
>
> Goldman (1999: 7)

But this conclusion seems way too hasty. Can dispensing with trans-mental processes really allow for a reliable knowledge of mathematical statements platonistically construed? Not quite, for as Sosa has pointed out:

> Some intra-mental processes lead reliably to beliefs about the physical objects in a subject's environment. Taking our experience at face value is normally such a process on the surface of the earth. Processes of reasoning or calculation also yield true beliefs reliably enough to make them warranted. Again, such processes for Goldman comprise no quasi-perceptual relations to the objects or facts that they enable us to know. So it is not by including such relations to abstracta that his processes of a priori warrant cause a problem; but they do cause a problem anyway. We can know and understand how it is that taking experience at face value is a reliable gateway to the shapes and colors of visible surfaces. But *this* explanation will apparently involve postulating perception of surfaces in appropriate conditions. By contrast, abstracta are imperceptible; so how could our intramental processes of reasoning or calculation put us reliably in touch with how it is with such abstracta? True, ethereal perception does not *constitute* the intramental processes. But it is still a mystery how these processes could be reliable about mind-transcendent facts without perception or some other causal mechanism to connect the two.
>
> Sosa (2003: 178–179)

If Sosa is right about this—and I believe he is—then we should not follow Goldman in thinking that his recent reliabilist MN can account for platonistic mathematical knowledge.

If what has been said so far is correct, the appeal to the notion of reliability discussed in mainstream epistemological debates will still make $\text{RBC}_1$ *ipso facto* unanswerable by the platonist, thus lending support again to the following modified versions of [*Incompatibility*$_C$] and [*Prejudiciality*$_C$]:

| | |
|---|---|
| [*Incompatibility*$_R$] | RTJ entails that platonism is unjustified. |
| [*Prejudiciality*$_R$] | Assuming RTJ in $\text{RBC}_1$ makes the latter prejudicial (if not question-begging) against the platonist. |

### 5.3.2 *Hale on CTK and Reliability*

As mentioned above, Hale's argument is addressed to CTK, not to RTJ. However, I think—and Hale himself suggests—that his argument can be split in two halves (see Hale 1987: Chap. 4). The first half addresses CTK, and its conclusions go in the same direction as [*Incompatibility*$_C$]. The second half is better seen as addressing theories of justification and deserves closer attention.

Hale wants to score a twofold result: that (a) on some strong reading, CTK precludes (platonist) mathematical knowledge but is itself untenable; and that (b) on some weak reading, CTK is compatible with (platonist) mathematical knowledge. He thus distinguishes between $\text{CTK}_S$ and $\text{CTK}_W$. Both

> will require, for knowledge that *p*, the existence of a causal connection of some sort between the state of affairs in virtue of which it is true that *p* and the putative knower's belief that *p*.
>
> Hale (1987: 93)

However, $\text{CTK}_\text{S}$ will involve "a requirement that the truth-conferring fact itself is suitably causally related to the putative knower's belief," while $\text{CTK}_\text{W}$ will "require (merely) that, for *x* to know that *p*, *x*'s grounds or evidence for *p* should be causally effective in producing his belief that *p*." Hale then makes two objections to $\text{CTK}_\text{S}$:

(a) $\text{CTK}_\text{S}$ seems untenable insofar as it also rules out knowledge of general empirical truths (e.g. "copper conducts electricity")
(b) even though adjustments can be made to $\text{CTK}_\text{S}$ in order to defuse point (a), the threat posed by $\text{CTK}_\text{S}$ to platonists adds nothing to threats posed by causal theories of reference.[27]

Either way, it is safe to take Hale as claiming that $\text{CTK}_\text{S}$ does rule out platonism (since it is either implausibly strong, or poses well-known and substantial threats).

When it comes to $\text{CTK}_\text{W}$, Hale notices that:

(c) it "can, and perhaps should, be viewed as elucidating the notion of justified belief, rather than replacing a JTB account by something else"

and that

(d) since it "involves no suggestion that the object of knowledge (i.e. what is known) must play a causal role, there is thus no reason to regard it as posing any particular threat to platonism."

As such, $\text{CTK}_\text{W}$ will be compatible with the possibility of (platonist) mathematical knowledge only if either the causal aspects involved in belief-forming processes allow cognitive processes to reliably form beliefs about states of affairs involving abstract objects, or there is some plausible understanding of reliability which involves no causal connection of any sort between the (facts constituting the) truth conditions of *p* and the belief that *p*, for some statement *p*.

However, as the previous section indicates, the first option is likely to be ruled out for most common readings of the notion of reliability. If this is so, appealing to $\text{CTK}_\text{W}$ will be equivalent to appealing to RTJ, and this, again, points towards [$Incompatibility_\text{R}$] and [$Prejudiciality_\text{R}$]. As regards the second option, I will later on suggest that any reading of reliability which obliterates causal connections reduces it to a far weaker notion of accuracy, and I will submit that versions of $\text{RBC}_1$ requiring that the platonist provides an explanation of accuracy, as opposed to reliability, of mathematical beliefs, pose no novel epistemological threat to platonism.

### 5.3.3 Conclusions to Section 5.3

Let us take stock. The following conclusions seem now available. On the assumption that Field, in formulating $\text{RBC}_1$, is surreptitiously appealing to notions of reliability available in the epistemological debate, we can argue: that Field implicitly assumes a causal explanation of reliability; that major accounts of reliability in the epistemological literature are either strongly causal in character (e.g. reliable-indicator theories), or relying on some causal understanding of reliable cognitive processes (process reliabilism); that $\text{RBC}_1$ should be read as asking the platonist for a justification of mathematical beliefs.

Like $\text{BC}_1$, $\text{RBC}_1$ is thus based on assumptions which immediately rule out platonism as incoherent—assumptions motivated by the unwarranted generalization of models of knowledge from the empirical to the mathematical.[28]

It turns out, then, that neither causal knowledge nor reliability (read: reliabilist justification) can legitimately be appealed to in a Benacerraf-style dilemma and in the corresponding challenge to the platonist, on pain of prejudiciality.

The obvious advice is to weaken the required constraints by appealing to weaker notions, so as to obtain a version of the challenge that would provide an arena where nominalists and platonists may dissent on an equal footing.

## 5.4 Weakening Constraints

What remains to be seen is thus whether there is any (interesting and non prejudicial) version of $\text{RBC}_1$ that can be offered with the help of weaker notions. Here, "weaker" stands for "less explicitly relying, either directly or indirectly, on causal notions." Theories of knowledge and theories of justification may be both relevant.

### 5.4.1 Weakening Reliability: Counterfactual Dependency

A first and obvious suggestion is to remain within the reliabilist camp, but to appeal to versions of reliabilism that do not require any causal connection between the knower and what is known. This would still satisfy Field's request that we should be able to explain why, if mathematicians believes that $p$, then $p$.

Counterfactual accounts of knowledge or justification are natural candidates, for these accounts would explain the required correlation while presumably remaining silent about what makes the correlation hold.

#### 5.4.1.1 Goldman's Relevant Alternative Theory

As previously mentioned, a form of counterfactual theory for perceptual knowledge was suggested in Goldman (1976). Goldman requires that a subject be ascribed knowledge that *p* "just in case he distinguishes or discriminates the truth of *p* from relevant alternatives" (Goldman 1976: 771).[29] This, however, is a modification of CTK by way of the introduction of further constraints concerning discriminating abilities. Insofar as it is a form of CTK, this theory falls prey to the same limitations considered above.

#### 5.4.1.2 Nozick: Tracking Truth

The best known counterfactual theory of knowledge has been suggested by Robert Nozick in Nozick (1981). According to Nozick, a subject knows that *p* iff the subject is in a position to track the truth of *p*. This means that, in order for a subject *S* to know that *p*, two counterfactual conditions must be added to those of truth and belief:

(i) if *p* were not true, the subject would not believe that *p* (the so-called sensitivity conditions, see fn. 22);
(ii) if *p* were true, the subject would believe that *p*.

This could be a good candidate for what we are looking for, since it is neither causal connections nor reliable processes which allow the subject to track the truth of *p*.

However, it is agreed by many that counterfactual theories of knowledge are inadequate for mathematical knowledge. If mathematical truths are necessary, the antecedent in condition (i) is necessarily false. Unless some appropriate account of counter-possibles is offered—and there seem to be no agreement on this—we lack a clear conception of what a counterfactual theory of knowledge would imply for mathematical knowledge.

Now, the preceding argument is in good standing only if mathematical truths are necessary. However, one might think that the existence of mathematical objects is contingent. On the nominalist side, Field has suggested this. On the platonist side, those arguing for platonism on the basis of (some particular) version of the indispensability argument could argue for the contingent existence of mathematical objects; Colyvan (2001) is a case in point. However, this view is controversial and certainly isn't part of the "standard view" that Benacerraf was addressing.

However, as Field claims (see Field 1989: 227–281), there may be a legitimate sense in which we speak of mathematical counter-possibles. After all, it makes perfect sense to wonder what would be the case if the Continuum Hypothesis or the Axiom of Choice were false. Generalizing, we might accept that it also makes sense to ask what would happen if basic arithmetical statements like "1 + 1=2" were false, however odd this may strike one at first. Maybe, thus, it is not implausible that some version of the counterfactual theory of knowledge be adopted in some

version of BC: call it $RBC_2$. $RBC_2$ would ask the platonist to explain how it is that mathematicians' mathematical beliefs are counterfactually dependent on mathematical facts. We have then two options:

(a) On the one hand, the platonist does indeed believe that mathematical counter-possibles are trivially true. $RBC_2$ would preclude (an explanation of) platonist mathematical knowledge from the start, since this kind of platonist would then clearly be unable to explain the required correlation between facts and beliefs.
(b) On the other hand, the platonist who doesn't hold principled views on counter-possibles will have to explain how some worldly (spatio-temporally located) facts—psychological states, i.e. mathematical beliefs—can be affected by what happens to be non-worldly (acausal, non spatio-temporally located) objects. Since the platonist takes mathematical objects to be causally inert and non-spatio-temporally located, however, this explanation will not be available.[30] Once again, $RBC_2$ rules out platonism much too easily.

### 5.4.2 *From Counterfactual Correlation to Accuracy*

A more radical route remains open. Let the proponent of $RBC_2$ grant the platonist that counterfactual dependence of beliefs on mathematical facts is unintelligible. Some sort of correlation still remains to be explained. Indeed, when Field introduced his revised challenge in Field [1988] (1989: 230–231), he did not state it in terms of reliability (as he did when he introduced it in Field (1989), as we saw above), but merely in terms of "regularity" between mathematical facts and mathematicians' beliefs: the fact that for most *p* it holds that if mathematicians believe *p*, then *p*. (Notice the use of "believe" as opposed to "believ*ed*.")

Thus, even when counterfactual dependency is dropped, we still have to explain actual correlation (see Field [1988] 1989: 238). Following recent discussion (see Liggins 2010; Linnebo 2006), I will call this *explanandum* "accuracy." Call, thus, $RBC_3$ the corresponding revised version of $BC_1$ in which the relevant explanandum is "accuracy." $RBC_3$ would then take the following form:

[$RBC_3$]

(i) [*Initial Plausibility*] We are (initially) justified in believing in the existence of mathematical objects as being those (abstract) objects our mathematical statements are about.
(ii) [*Accuracy Claim*] Mathematicians' mathematical beliefs are accurate.
(iii) [*Explanation*] The platonist must explain [*Accuracy Claim*].
(iv) [$Impossibility_A$] It is not possible for the platonist to explain [*Accuracy Claim*].
(v) [*Defeat*] [$Impossibility_A$] defeats [*Initial Plausibility*].

Accuracy seems the most neutral *explanandum* we can ask for. So our question now should be: is $\text{RBC}_3$ once more guilty of requiring the impossible from the platonist? Is this revised version of $\text{BC}_1$ once again so strong that the very possibility of platonism turns out to be barred by the very nature of the *explanandum*? The answer, this time, seems finally to be in the negative. There is *nothing*, in the notion of accuracy, i.e. in the very fact that mathematicians' beliefs are true most of the time, to prevent a platonistic explanation.[31]

But maybe this is too quick. For one might suspect—as Field does, immediately after suggesting $\text{RBC}_3$—that the acausal or abstract character of the mathematical entities allegedly constituting the facts conferring truth on our beliefs is an obstacle to the explanation of accuracy. But assuming this entails either (i) committing a now familiar fallacy of [*Prejudiciality*]; or (ii) begging the question against the platonist in yet a different sense. It entails (i) if the suspicion about the relevance of the acausal or abstract character in the explanation of accuracy is motivated by the generalization of modes of knowledge that are appropriate in the empirical case to the non-empirical case. And this is a case of hasty generalization we should reject. It entails (ii) because there is no indication that the acausal or abstract character of any entity should prevent an explanation of the accuracy of our beliefs.

Accordingly, this may not be the sought-for middle position where the challenge to the platonist is *both* non-prejudicial *and* substantial. For we seem to have immediately jumped to the opposite extreme where it is not even clear that *any* genuine challenge is proposed. To be more precise: there seems to be no challenge to the platonist in addition to the familiar and innocent request that a philosophical account of an area of discourse be supported by good arguments and evidence. But this very general challenge makes mathematical platonism no worse off than any other philosophical view whatsoever.[32] Hale and Wright rightly emphasize that:

> […] the general issue of reconciling semantics and epistemology for mathematics is not just a challenge for would-be platonists—it faces all philosophical positions which allow that pure mathematics presents, as it is normally taken to do, a substantial proper part of human knowledge.
>
> Hale and Wright (2002: 103)

But now, how and when should this generalization stop? Why shouldn't the challenge be simply that of reconciling, with no additional constraint, semantics and epistemology generally, *for all philosophical positions whatsoever*?

## 5.5 A Dilemma's Dilemma

The preceding discussion leaves several questions unanswered. First of all, despite the initial impact of Benacerraf's original challenge and of Field's revised version, it isn't clear anymore that a coherent epistemological challenge to platonism along the lines suggested by Benacerraf's dilemma is still available.

I am not denying that mathematical platonism, in whatever version, has the heavy burden of giving an account of mathematical knowledge, of our epistemic access to mathematical objects, and of our methods of justification for mathematical beliefs. What I want to suggest is that it isn't clear that the structure of Benacerraf's original argument, or of Field's revised version, really adds anything to the general and obvious request that platonists supplement their view with a plausible and reasonable epistemology.

I think it is safe to assume the validity of a familiar and generally accepted form of objection that we might call the [*No Additional Charge*] objection (or, better still, counter-objection): an argument $m$, construed as an objection to some position $x$, poses a significant challenge to $x$ only insofar as it suggests that $x$ has other difficulties to deal with in addition to those that other objections $n_1$, $n_2$,… have already suggested.

> [*No Additional Charge*]
>
> is clearly more effective when other or previous challenges to $x$ point to a problem that $x$ shares with other philosophical views, either in the same vicinity or perhaps in other areas of inquiry.

The dialectic situation of Benacerraf's and Field's epistemological challenges to platonism suggests a whole family of dilemmas, varying along two different theoretical axes. On the one hand, they vary according to how demanding the semantic and epistemological notions involved in their formulation are. Call this the *robustness* axis. On the other hand, they vary according to how specifically a particular area of inquiry is targeted by the challenge. Call this the *specificity* axis.

I have been suggesting that a sensible challenge to platonism, along the lines of a Benacerraf-style dilemma, may be obtained only if a suitable point of intersection is secured for these two axes: in other words, only in case one or more dilemmas may be found that: (i) rely on notions *weak enough* so that platonism doesn't turn out to be a non-starter, i.e. so that the corresponding challenge to the platonist isn't prejudicial; and (ii) deliver a *self-standing* challenge to philosophical views about *mathematics* (be they platonism or any other view), i.e. a challenge that is both new with respect to other available challenges to philosophical accounts of mathematics, and *specific to mathematics*.

The foregoing discussion, however, seems to suggest that no such fruitful point of intersection will be forthcoming. We already saw that:

[$\text{BD}_1$]

(a) A referential (i.e. Tarskian) account of mathematical truth must be compatible with a causal account of mathematical knowledge
(b) However, the accounts are mutually exclusive

fails with respect to the robustness axis. The notion involved on the epistemological side makes the corresponding challenge $\text{BC}_1$ prejudicial (if not question-begging) against its opponent.

The Benacerraf-style dilemma underlying the discussion in Sect. 5.3 could be stated as follows:

[$BD_2$]

(a) A referential (i.e. Tarskian) account of mathematical truth must be compatible with a reliabilist account of justification for mathematical beliefs.
(b) However, …

In this case, the notion one resorts to for the epistemic horn of the dilemma is weaker than the one resorts to in $BD_1$. However, we have seen that there is a good deal of evidence that $BD_2$ also fails with respect to the robustness axis, since it points once again towards corresponding versions of [*Prejudiciality*].

We may go on along the robustness axis, by gradually weakening the notions involved. In order to avoid [*Prejudiciality*] (or begging the question), we are bound to arrive at what we found to be the most general notion available, i.e. accuracy:

[$BD_3$]

(a) A referential (i.e. Tarskian) account of mathematical truth must be compatible with an account of the accuracy of our mathematical beliefs.
(b) However, …

It now seems that we are immune from the threat of prejudiciality along the robustness axis. However, different threats loom large, for now the corresponding challenge to the platonist—$RBC_3$—reduces to a demand for an explanation of the correlation between belief(s) and fact(s) that is not in any clear sense motivated by qualms about the role that abstract mathematical objects would play in constituting the relevant truth-conferring facts. But, in this case, $RBC_3$ fails with respect to the other axis, the specificity axis, and falls prey to the [*No Additional Charge*] counter-objection: no additional challenge is addressed to the platonist other than challenges that are already independently available, such as the general request that *any* philosophical view about a given area of discourse accommodates an acceptable epistemology for that area of discourse.

From $BD_3$, we can proceed through progressive generalizations of both requirements (the semantical and the epistemological) in the dilemma, so as to obtain more and more comprehensive versions:

[$BD_4$]

(a) A suitable/acceptable/good account of mathematical truth must be compatible with a suitable/ acceptable/good account of the epistemology for mathematics.
(b) However, …

[$BD_5$]

(a) An account of mathematical truth must be compatible with an account of the epistemology for mathematics.
(b) However, …

[$BD_6$]

(a) An account of truth for an area of discourse *A* must be compatible with the epistemology for *A*.
(b) However, …

Needless to say, each of these dilemmas fails, in increasing order, with respect to the specificity axis. $\text{BD}_4$ reduces to the obvious demand for a suitable epistemology of mathematics that isn't even related to any particular semantics for mathematical discourse. $\text{BD}_5$ still generalizes this very general challenge. Last and least, $\text{BD}_6$ does not even include any reference to mathematics.

Arguably, in these last three formulations, nothing even guarantees that clause (b) meaningfully provides *any dilemma at all*. Nothing seems, for instance, to support with the required generality (unless additional assumptions are made) that if one can provide a suitable semantics for mathematics (whatever that may be), one cannot at the same time provide a suitable epistemology for it, as $\text{BD}_5$ requires. Things are even worse with $\text{BD}_6$, for in that case the alleged dilemma brought about by (b) would have the radically skeptical consequence that *no area of discourse whatsoever* could be given an acceptable philosophical account.

If what has been said so far is correct, epistemological challenges to platonism modelled on Benacerraf-style dilemmas face the following situation. There are two desirable requirements that any such dilemma should satisfy. On the one hand, it should not rely on notions so robust as to make the corresponding challenge to the platonist prejudicial, let alone question-begging. On the other, it should not be so general that no new or dedicated threat is raised against mathematical platonism, or at least against philosophical accounts of mathematics generally. If the evidence gathered so far (as incomplete as it may be) is correct, it points to a now very familiar Benacerraf-style challenge. In Benacerraf's own words: "these then are the two requirements. Separately, they seem innocuous enough" (Benacerraf 1973: 668). Both are desirable, and each can be argued for. But any dilemma "can be identified with serving one or another of these masters *at the expense of the other*" (Benacerraf 1973, 661).

## Notes

1. For discussions of epistemological challenges to platonism, see Linnebo (2006), Liggins (2010).
   For a survey of recent responses to Benacerraf's dilemma, see Panza and Sereni (2013).
2. At least since Burgess and Rosen (1997).
3. Of course, as Benacerraf (1973) discusses, formalists will have to meet their own version of the challenge.
4. See Maddy (1990a, b).
5. In what follows, unless crucial to the argument and specified, whether what is at stake is knowledge *that* (mathematical facts holds) or knowledge *of* (mathematical objects) should be clear from the context.
6. A platonist position according to which all mathematical truths are unknowable is therefore considered untenable.
7. Depending on interpretations of the text, Benacerraf may be taken to merely assert that CTK should apply to mathematical statements because it is the best theory currently available, and not because a single uniform account of our

knowledge in radically diverse areas is required. The latter, however, seems to be the standard interpretation.

8. The anti-platonist will argue that CTK is the best theory of knowledge available for empirical truths and will recommend an extension to the case of mathematical truths. The anti-nominalist will deem this extension unjustified, for nothing garantees that the two areas can or indeed should be treated as epistemically on a par. The dispute will lead to a reciprocal burden-of-proof argument with no clear way out.
9. See Cheyne (2001) for a strenuous criticism of (a).
10. The same holds for one of the arguments presented in Steiner (1973: 59–60). Steiner argues that the following fully general formulation of CTK:

    (CTK*) One cannot know that *p* unless the fact that *p* causes one's knowledge (or belief) that *p*

    presupposes the existence of facts. Facts are abstract entities and, as such, cannot enter in causal interactions. So (CTK*) precludes all knowledge since it states that knowledge is possible just in case some fact produces the belief that *p*. This is clearly consistent with [$Incompatibility_C$] and [$Prejudiciability_C$].

11. There is another possible reply, i.e. that the (only) essential feature of causation is that causal relations support appropriately specified counterfactuals, but Callard does not want to go this way because counterfactual analyses of knowledge seem inadequate for knowledge of necessary truths (as we are assuming mathematical truths to be). More on this point below.
12. On a more relaxed reading of "effecting changes," we sometimes speak of the effecting of a change in connection with non-causal relations, as in, e.g., "My love for philosophy changed my life." But on this reading the effecting of the change can at best be a necessary condition for causation. It doesn't provide a definition of it.
13. It is important to notice that a full defence of [$Impossibility_R$] will depend on the sort of explanation that a platonist is required or allowed to offer. See Linnebo (2006) for an elaborate discussion on this point.
14. As Field himself acknowledges, this alone does not answer [*Explanation*]. According to Burgess and Rosen, rather than proper correlation, we might in the end need the explanation of a mere conjunction. Since the reliability of our beliefs in theorems is warranted by deduction, and since classical mathematics can be reduced to set theory and all axioms of set theory can be summed up in the claim "The full hierarchy of sets exists," what we need is just an explanation of the conjunction "The full hierarchy of sets exists and it is believed [i.e. mathematicians believe] that the full hierarchy of sets exists" (Burgess and Rosen 1997: 45).
15. Consider the following, stronger formulation of Field's revised challenge:

(i*) [*Semantic Justification*] A referential (Tarskian) account of truth entails that the mathematical terms occurring in our truly believed mathematical statements have referents and that such referents exist.
(ii*) [*Knowledge Claim*] Mathematicians mostly causally know the mathematical statements they believe.
(iii*) [*Explanation*$_{\mathrm{K}}$] The platonist must explain [*Knowledge Claim*].
(iv*) [*Impossibility*$_{\mathrm{K}}$] It is not possible for the platonist to explain [*Knowledge Claim*].
(v*) [*Conflict*] [*Impossibility*$_{\mathrm{K}}$] is in conflict with [*Semantic Justification*].

Premise (ii*) may strike one as harsh. However, this is consistent with the stringent formulation of BD$_1$, so that the argument just given turns out to be a reformulation of BC$_1$. An alternative version would make the causal analysis of knowledge explicit as an additional premise—but this would disrupt the similarity in structure with RBC$_1$.

16. Linnebo directly formulates RBC$_1$ in terms of justification (see Linnebo 2009: 18; emphasis mine):

    1. Mathematicians are reliable, in the sense that for almost every mathematical sentence *S*, if mathematicians accept *S*, then *S* is true.
    2. For belief in mathematics to be *justified*, it must at least in principle be possible to explain the reliability described in Premise 1.
    3. If mathematical platonism is true, then this reliability cannot be explained even in principle.

    If these three premises are correct, it will follow that mathematical platonism undercuts our *justification* for believing in mathematics.
    Notice that "belief in mathematics" in Premise 2 should not (not only?) be taken as referring to our philosophical justification for believing mathematical platonism, as opposed to mathematicians' justification for their belief. The following claim by Linnebo makes this clear (Linnebo 2009: 19; emphasis mine):

    > Premise 2 seems fairly secure. If the reliability of some belief formation procedure could not even in principle be explained, then the procedure would seem to work purely by chance, thus undercutting any justification we have *for the beliefs produced in this way*.

17. We are granting that ascertaining the truth of a mathematical belief in order to assess reliability, and hence justification and possibly knowledge, is a non-circular process. But is this so? Ascertaining the truth of a mathematical statement *p* is tantamount to know that *p*: clearly, *p* can be true without me knowing that *p*, but how can I *ascertain* that *p* without thereby coming to know that *p*? If "ascertaining truth" is to be relevant in the explanation of reliability, it *must* entail knowledge that *p*, or at least justified belief that *p*.
    But this goes in my direction. For now RBC$_1$ would again be requiring of the platonist something that she is in principle unable to offer—an explanation of how she can have reliable beliefs that *p* that itself presupposes some account of

how she can come to know (and hence have reliably-formed justified beliefs) that *p*. This is either an inconsistent request in general, or (should it be accepted in cases of empirical knowledge) one which is in principle inaccessible to the platonist and therefore makes $\text{RBC}_1$ too strong.
On a very different reading, the request could be seen as one of pointing at some indirect grounds for the truth of some mathematical statement p. We might have a genuine challenge here, which nonetheless can be tackled in several ways (like those advanced by neo-logicism and indispensability arguments), and it is unrealistic to think that this is what $\text{RBC}_1$ is really asking for.

18. Unless, of course, one has an infallibilist notion of justification, for which even high probability of true outputs would not suffice. Notice that we are not discriminating here between externalist and internalist views: tests for internal coherence may well be considered among the relevant candidate processes. Notice also that on a strongly externalist reading, we may be justified in believing *p*, granted that *p* is reliably formed, even if we have no explanation at all of the mechanism that has produced *p*. This, however, does nothing to defuse the point that if we have such an explanation, then our belief will count as justified. In any case, strongly externalist readings seem inadequate for an account of mathematical justification and knowledge.
19. Mind-independence is the target of Divers and Miller response-dependent account of mathematics, so it does play a special role in their exposition.
20. Wright (1983) was published six years earlier and Field already discussed it in Field (1984). Burgess and Rosen (1997: 42) suggest that Field's challenge is neither addressed to Maddy's platonism nor to neo-logicist platonism, since Field argues against these in separate places. But it is unclear why Field's $\text{RBC}_1$ should not also figure among the neo-logicists' concerns, as the original $\text{BC}_1$ indeed does.
21. More specifically, see Armstrong (1973: 170):
A's non-inferential belief that c is a J is a case of non-inferential knowledge if, and only if:

    (i) *Jc*
    (ii) (∃H)[Ha & there is a law-like connection in nature (*x*)(*y*) {if H*x*, then (if B*x*J*y*, then J*y*)},

    where "*x*" ranges over "beings capable of cognition," and "B*x*J*y*" means "*x* believes that *y* is a J."
22. The example goes: travelling in the countryside, Henry looks at what, unbeknownst to him, is the only real barn in a neighborhood of fake barns, and forms the belief (by perception, in normal conditions) "That's a barn." The belief is actually true and appropriately causally formed. However, we would not attribute knowledge to Henry, for he could easily have been mistaken in close words in which that belief is false. In contemporary epistemological jargon, Henry's belief violates a constraint of sensitivity. A belief that *p* is sensitive if, in all closest possible worlds in which *p* is not true, *S* would not

believe that *p*. In the close possible world in which Henry looks at a nearby fake barn, his belief is false but he would still have it. For a discussion of sensitivity and related constraints on beliefs, cf. Pritchard 2008.

23. Goldman originally suggested that our everyday conception of justification is vague enough in this respect, so that different interpretations may be appropriate in different circumstances (i.e. in assessing different processes), although he generally leaned towards a propensity interpretation in his later work.
24. For a discussion of Casullo's claims, see Cheyne (2001: 126–130). Compare with what Goldman and Pust say about the role of intuitions in securing knowledge of universals:

> The chief difficulty for this approach comes with the assumption that intuition is a basic evidential source, a source of information about universals. Is there any reason to suppose that intuitions could be reliable indicators of a universal's positive and negative instances (even under favorable circumstances)? The problem is the apparent "distance" or "remoteness" between intuitions, which are dated mental states, and a nonphysical, extra-mental, extra-temporal entity. How could the former be reliable indicators of the properties of the latter? This is similar to the problem Benacerraf (1973) raises about the prospects for mathematical knowledge on any [p]latonistic view of mathematics. Benacerraf, however, assumes that a causal connection with the object known is necessary for knowledge. We deliberately have not imposed such a requirement for a basic evidential source. Nor have we imposed the requirement of a counterfactual dependence between states of affairs that make an intuition's content true and the occurrence of such intuitions. We have only imposed the reliable indicatorship constraint. However […] wherever it is obscure, as it is here, how a causal relation or counterfactual dependence of the right sort could obtain, there are grounds for serious doubt that the reliable indicatorship relation obtains. Some philosophers […] might reply that abstractness per se does not exclude causal relations. Nonetheless, we certainly lack any convincing or even plausible story of how intuitions could be reliable indicators of facts concerning universals.
>
> Goldman and Pust (2002: 80)

25. Goldman defines Moderate Naturalism as the conjunction of the following two claims:
    (a) An epistemic warrant or justification is a function of the psychological (perhaps computational) processes that produce or preserve belief.
    (b) The epistemological enterprise needs appropriate help from science, especially the science of the mind.
26. Goldman (Goldman 1999) considers four properties that are traditionally associated with a priori knowledge or (as he calls it here) warrant: (1) a non-experiential, i.e. non-perceptual, source or basis, (2) necessity, (3) a subject-matter of abstract, eternal objects, (4) infallibility, (5) certainty, and (6) rational unrevisability (incorrigibility). He denies that (4), (5) and (6) should be taken as features of the a priori. This might be questionable, but it is not relevant here. More relevant is that he sets aside (3), claiming the following:

> Here I want to stick to my earlier resolve to stick to the *epistemological* questions concerning the a priori and avoid the *metaphysical* questions. Thus, I want to remain neutral on the issue of what the subject-matter of the a priori has to be. To be more precise, although I am willing to concede that only beliefs on certain topics or in certain domains will qualify

as warranted a priori, I want to remain neutral on the question of what the *truth-makers* are in those domains. I want to be able to concede the possibility of a priori warrant about arithmetic without taking a position on what numbers are or must be. Given this desire for metaphysical neutrality, it is obviously unacceptable to make an abstract subject-matter a necessary condition for a priority.

27. Cheyne (2001: Chap. 7) reviews and criticize Hale's argument, at least insofar as existential knowledge is involved.
28. Kasa (2010) has independently suggested that Field's revised challenge is even worse off than Benacerraf's original one. Kasa argues that Field's version of the challenge entails that every valid explanation of (a case of) knowledge must assume X, where X is something inaccessible to the platonist. According to Kasa, the best candidate for X is either that the entities involved in the truth-conferring facts are causally efficacious or that they are spatio-temporally located. Kasa's argument thus suggests that [*Causality*] can be generalized into:

[*Non-Abstractness*] Non-abstractness constraints on either knowledge or justification rule out direct knowledge of abstract objects, knowledge of truths involving abstract objects, and justification of truths involving abstract objects.

If we add Kasa's conclusion to ours, it follows that $\text{RBC}_1$ is prejudicial insofar as it assumes not only causal constraints on knowledge, but also non-abstractness constraints on justification.

29. E.g., from the truth of "There is a barn in front of me" and the truth of "There is a fake barn in front of me and all other similar barns in the vicinity are real"; but not from the former and the truth of "There is no external world."
30. But see Linnebo (2006) for a proposal on how to make sense of such dependency.
31. Notice that, in recent proposals by Sosa (2007, 2011) where accuracy does play a role, not only is accuracy insufficient for second order justification of one's own first order beliefs (what Sosa calls "reflective knowledge"), it isn't even sufficient for first order justification or knowledge (what Sosa calls "animal knowledge"). Beyond accuracy (the property of a performance, e.g. such as believing, of being successful), what is needed for animal knowledge is also adroitness (the property of a performance of being accurate because adroit, i.e. accurate insofar as it is the effect of the manifestation of a competence and not the outcome of luck or accident). In order to have reflective knowledge, what is needed is an aptly possessed second order belief concerning the aptness of one's first order beliefs. This is way more than what $\text{RBC}_3$ would be asking of the platonist.
32. It becomes indeed the very general challenge that Christopher Peacocke calls "the Integration Challenge," i.e. the very general challenge, proper to such diverse areas such as discourse about mathematics, the past, necessity, self-knowledge, etc., of showing how we can "reconcile a plausible account of what is involved in the truth of statements of a given kind with a credible

account of how we can know those statements, when we do know them" (Peacocke 1999: 1).

**Acknowledgments** Earlier versions of this paper were presented at the Paul Benacerraf Workshop (May 10-11, 2012, Paris, Collège de France/IHPST); at the European Epistemology Network Conference (Universities of Bologna and Modena & Reggio Emilia, June 28-30, 2012); at the ECAP8 (28.08–02.09, 2014, Bucharest). Many thanks to the audiences of those conferences for helpful comments. Special thanks go to Øystein Linnebo, Marco Panza, and Eva Picardi for fruitful discussions which led to several improvements.

## References

Armstrong, D. M. (1973). *Belief, truth and knowledge*. Cambridge: Cambridge UP.

Azzouni, J. (2008). A cause for concern: Standard abstracta and causation. *Philosophia Mathematica, 16*(3), 397–401.

Benacerraf, P. (1973). Mathematical truth. *Journal of Philosophy ,70*(19), 661–679.

Benacerraf, P., & Putnam, H. [1964] (1983). Philosophy of mathematics: Selected readings - 2nd edition (1st edition 1964). Cambridge: Cambridge University Press.

Burgess, J. P., & Rosen, G. (1997). *A subject with no object: Strategies for nominalistic interpretation of mathematics*. Oxford: Oxford UP.

Callard, B. (2007). The conceivability of platonism. *Philosophia Mathematica, 15*(3), 347–356.

Casullo, A. (1992). Causality, reliabilism, and mathematical knowledge. *Philosophy and Phenomenological Research, 52*(3), 557–584.

Cheyne, C. (2001). *Knowledge, cause and abstract objects: Causal objections to platonism*. Dordrecht: Kluwer Academic Publishers.

Colyvan, M. (2001). *The indispensability of mathematics*. Oxford: Oxford UP.

Divers, J., & Miller, A. (1999). Arithmetical platonism: Reliability and judgement-dependence. *Philosophical Studies, 95*(3), 277–310.

Field, H. H. (1984). Platonism for cheap? Crispin Wright on Frege's context principle. *Canadian Journal of Philosophy, 14*, 637–662.

Field, H. H. (1989). *Realism, mathematics, and modality*. Oxford: Basil Blackwell.

Field, H. H. [1988] (1989). *Realism, mathematics and modality* (pp. 227–281), originally published as Realism, mathematics and modality. *Philosophical Topics, 16*(1), 1988, 57–107.

Goldman, A. I. (1967). A causal theory of knowing. *The Journal of Philosophy, 64*(12), 357–372.

Goldman, A. I. (1975). Innate knowledge. In S. P. Stich (Ed.), *Innate ideas* (pp. 111–120). Berkeley: University of California Press.

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy, 73* (20), 771–791.

Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.) *Justification and knowledge* (Vol. 17, pp. 1–23), Dordrecht: D. Reidel Publishing Company.

Goldman, A. I. (1986). *Epistemology and cognition*. Cambridge, Mass: Harvard UP.

Goldman, A. I. (1999). A priori warrant and naturalistic epistemology: The seventh philosophical perspectives lecture," *Noûs*, *Supplement: Philosophical Perspectives*, *33*(s13), 1–28.

Goldman, A. I. (2011). Reliabilism. In E. N. Zalta, (Ed.), *Stanford encyclopedia of philosophy* (Spring 2011 Ed.). http://plato.stanford.edu/archives/spr2011/entries/reliabilism/

Goldman, A. I. & Pust, J. (2002). Philosophical theory and intuitional evidence, In A. Goldman (Ed.), *Pathways to knowledge* (pp. 73–94 ), Oxford and New York: Oxford UP. (Originally published as chapter 11 of *Rethinking Intuition: The Psychology of Intuition and its Role in*

*Philosophical Inquiry* (pp. 179–198), M. Depaul & W. Ramsey (Eds.) (1998), New York; Rowman and Littlefield.
Hale, B. (1987). *Abstract objects*. Oxford: Oxford UP.
Hale, B., & Wright, C. (2002). Benacerraf's dilemma revisited. *European Journal of Philosophy, 10*(1), 101–129.
Hart, W. D. (1977). Book review of Mark Steiner's "Mathematical Knowledge". *The Journal of Philosophy, 74*(2), 118–129.
Kasa, I. (2010). On Field's epistemological argument against platonism. *Studia Logica, 96*(2), 141–147.
Liggins, D. (2006). Is there a good epistemological argument against platonism? *Analysis, 66* (290), 135–141.
Liggins, D. (2010). Epistemological objections to platonism. *Philosophy Compass, 5*(1), 67–77.
Linnebo, Ø. (2006). Epistemological challenges to mathematical platonism. *Philosophical Studies, 129*(3), 545–574.
Linnebo, Ø. (2009). Platonism in the philosophy of mathematics. In E. Zalta, (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2013 Ed.). http://plato.stanford.edu/archives/win2013/entries/platonism-mathematics/
Maddy, P. (1990a). Physicalistic platonism. In A. Irvine (Ed.), *Physicalism in mathematics* (pp. 259–289). Dordrecht: Kluwer Academic Publishers.
Maddy, P. (1990b). *Realism in mathematics*. Oxford: Oxford UP.
Nozick, R. (1981). *Philosophical explanations*. Cambridge, Mass: Harvard UP.
Panza, M. & Sereni, A. (2013). *Plato's problem. An introduction to mathematical platonism*. Basingstoke: Palgrave Macmillan.
Peacocke, C. (1999). *Being known*. Oxford: Oxford UP.
Pritchard, D. (2008). Sensitivity, safety, and anti-luck epistemology. In John Greco (ed.), The Oxford Handbook of Skepticism. Oxford: Oxford UP.
Ramsey, F. P. (1931). Knowledge. In R. B. Braithwaite (Ed.) *The foundations of mathematics and other logical essays*, pp. 258–259. New York: Harcourt, Brace and Company.
Sosa, E. (2003). Ontology, understanding and the a priori. *Ratio, 16*(2), 178–188.
Sosa, E. (2007). *A virtue epistemology*. Oxford: Oxford UP.
Sosa, E. (2011). *Knowing full well*. Princeton, New Jersey: Princeton UP.
Steiner, M. (1973). Platonism and the causal theory of knowledge. *The Journal of Philosophy, 70* (3), 57–66.
Strawson, P. F. (1979). Universals. *Midwest Studies in Philosophy, 4*(1), 3–10.
Unger, P. (1968). An analysis of factual knowledge. *The Journal of Philosophy, 65*(6), 157–170.
Wagner S. J. (1996) Prospects for platonism. In A. Morton & S. P. Stich (Eds.), *Benacerraf and his critics* (pp. 73–99).
Wright, C. (1983). *Frege's conception of numbers as objects*. Aberdeen: Aberdeen UP.

# Part II
# Logicism, Fictionalism and Structuralism

[R]eshuffling the *designata* of our number words without making a corresponding readjustment in their connection to our practices of counting, etc. is perilous business and can lead to grievous error if remedial measures are not taken.

Benacerraf (1996b: 48)

# Chapter 6
# Benacerraf on Logicism

**Sébastien Gandon**

I will be making three distinct but interrelated claims in this paper. First, I will attempt to show that Benacerraf (1981) ("Frege: The Last Logicist") should be seen as a distant continuation of the better known Benacerraf (1965) ("What Numbers Could Not Be") and Benacerraf (1973) ("Mathematical Truth"). Second, I will claim that logicism does not fit in with the dilemma that Benacerraf set forth in 1973. It seems to me that the relative neglect of Frege in 1973 is the source of Benacerraf's wish to come back to Frege in 1981. Third, I will suggest that the way in which Benacerraf saved Frege in 1981 should have some impact on his 1973 claim that philosophers must take mathematical discourse at face value—and, more generally, on the way Benacerraf contrasts philosophical problems with mathematical ones.

I will develop these ideas in five separate sections. In the first, I will present a summary of Benacerraf (1981)'s main points. In the second, I will draw connections between the 1981 paper and the 1965 and 1973 papers. In the third, I will continue to examine Benacerraf (1981) and to focus on its main thesis, namely that Frege's project is neither philosophical nor epistemological, but mathematical. In Sect. 6.4, I will contrast Benacerraf's reading of Frege with the neo-logicist program. In Sect. 6.5, I will challenge the relevance of Benacerraf's dilemma by radicalizing Benacerraf's 1981 interpretation of Frege.

S. Gandon (✉)
PHIER, Université Clermont, Auvergne, France
e-mail: sgandon0@gmail.com



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_6

## 6.1 Frege and neo-positivist Logicism

The main goal of Benacerraf 1981 is to show that Frege's project hardly bears a likeliness to the program the neo-positivists called "logicism." What is neo-positivist logicism? It is an epistemological thesis about the place of mathematical knowledge within knowledge in general:

> The philosophical point of advancing the view was nakedly epistemological: logicism, if it could be established, would show how our knowledge of mathematics could be accounted for by whatever would account for our knowledge of language.
>
> Benacerraf (1981: 18)

According to Benacerraf, the positivist conceptual framework is quite unrefined. The basic idea is that there are two kinds of knowledge, analytic knowledge and synthetic knowledge. Roughly said, the former derives from our knowledge of language and the meaning of our words, while the latter derives from our interaction with the world. For the positivists, an analytically true statement is a statement which is true in virtue of the meanings of the words occurring in it. Now, if mathematics is nothing else than logic, and if logical knowledge derives entirely from a knowledge of language, then mathematical knowledge is analytic, and can easily find its place in an empiricist (i.e. anti-Kantian) epistemology. Let me quote a telling passage where Benacerraf comments Hempel's typical approach:

> A striking example is the position advanced by C. G. Hempel in an article that, although breaking no really new ground, presented a nuclear point of view as only Hempel can. According to Hempel, the Frege-Russell definitions of number, 0, successor, and related concepts have shown the propositions of arithmetic to be analytic because they follow by stipulative definitions from logical principles. What Hempel has in mind here is clearly that in a constructed formal system of logic […], one may introduce by stipulative definition the expressions "Number," "Zero," "Successor" in such a way that sentences of such a formal system using these introduced abbreviations and which are formally the same as […] certain sentences of arithmetic […] appear as theorems of the system. He concludes from that undeniable fact that these definitions show the theorems of arithmetic to be mere notational extensions of theorems of logic, and thus analytic.
>
> Benacerraf op. cit.: 20

According to Benacerraf, the positivist view is untenable. It implies that logical knowledge itself is analytic (logical propositions being true in virtue of their meaning). And this would mean that the logical theorems of, e.g., first order logic, get their truth in virtue of the implicit definitions of the logical constants. Benacerraf refers here to Quine's argument, set forth in "Truth by Convention" (Quine [1935] 1976), according to which the very application of the rules which implicitly define the logical constants requires the use of logic. The positivist solution is circular:

> The only solution that seems to be offered in such a case [i.e. when logic must comprise enough set theory (or suitable equivalent) to yield enough mathematics] is that the axioms constitute an implicit definition of the concepts. This is a form of conventionalism that construes the axioms as stipulations that are to govern the use of the terms they contain. […] As an explanation of how sentences of logic in fact get their truth-values it is worthless, as Quine and others have made abundantly clear.
>
> Benacerraf op. cit.: 20–21

But this critical stance is not the main tenet of Benacerraf's argument. His most important claim is that Frege never shared this mistaken view of logical and mathematical truth. According to Benacerraf, Frege ascribed to Kant a truth-in-virtue-of-meaning conception of analytic judgments. And Frege criticized Kant on this point: the target of Frege's distinction between the content of a judgment and its justification is precisely Kant's view that analytic judgments are analytic in virtue of their content.[1] Benacerraf rightly concludes that the criticism Frege launched against Kant could also be launched against the neo-positivist logicists:

> For Kant, the distinction between analytic and synthetic propositions was primarily a distinction in the content of the propositions. And the epistemological point was that this distinction in content had, for the analytic propositions, the immediate consequence that they […] were knowable independently of experience just on the basis of a consideration of their content. […] Twentieth-century "logicists," following Kant in this respect, accorded a priori status to an enlarged class of analytic propositions on the basis of their content — for truth-in-virtue-of-meanings is simply an extension of Kant's distinction and of the epistemological analysis that went along with it.
>
> Benacerraf op. cit.: 25

According to Benacerraf, one should, in order to understand Frege, distinguish two views of analyticity: the somewhat vague view that analytic truths are true in virtue of their meaning, and the more precise view that an analytic truth is a truth which can be transformed into a logical truth by way of definitions. Frege—and Quine as well—defends the second conception. But this last characterization bears little ostensible relation to the truth-in-virtue of meaning definition of analyticity—a definition which confers its epistemological importance on the logicist project. Frege's logicist undertaking (just like Quine's) cannot then be easily enrolled in the epistemological program that underlay the neo-positivist approach.

In brief, Benacerraf (1981) (just like Musgrave 1977) denounces a kind of cheating. When the positivists assert that mathematics is logic, they usually rely on the technical sense of analyticity (i.e. on the one that Frege and Quine used: an analytic truth is a truth which can be transformed into a logical truth by way of definitions). But when they want to account for the philosophical import of logicism, they resort to the truth-in-virtue-of-meaning sense, because only it "bore the epistemological burden of persuading us that analytic propositions were also a priori." Positivists conflated two theses that should indeed be kept separate:

> If, as W. V. [O.] Quine has done, one defines analytic truth as transformability into a logical truth by meaning-preserving definitions, it becomes a trivial matter that the laws of logic are analytic; but such a definition, as applied to logic, bears little ostensible relation to the traditional account of analyticity as truth-in-virtue-of-meanings. Yet it was this latter explanation which bore the epistemological burden of persuading us that analytic propositions were also a priori.
>
> Benacerraf op. cit.: 19

According to Benacerraf, Frege never claimed that mathematical knowledge derived from a knowledge of language. Frege thus never had an epistemological goal. What was then the philosophical pay-off of Frege's program?

## 6.2 Frege's Logicism: an epistemological or a semantical program?

Before presenting Benacerraf's answer, let me stress how substantial the problem is. A paragraph of the 1981 article reveals that, even though he was aware of the positivists' confusion, Benacerraf still had trouble untangling the neo-positivist epistemological view of logicism from the Quinean-Fregean "semantical" view:

> The view I have been calling "logicism" is evidently an amalgam of two views: a semantical thesis to the effect that arithmetic is a definitional extension of logic (Frege's view) and an epistemological claim about how this explains the a priori character of arithmetic (the positivists' view). Evidently, one can […] reserve the title for the semantical thesis alone, in which case Frege was certainly as much of a logicist as his followers […]. I chose the present method partly for dramatic effect and *partly because I am not really sure how clearly the two theses can be untangled from one another — how much the philosophical motivation behind a given form of the semantical thesis infects the thesis itself* [emphasis mine].
>
> Benacerraf op. cit.: 35

To understand Benacerraf's trouble, one must go back to 1973. Recall that in Benacerraf (1973), Benacerraf set forth two desiderata which any acceptable philosophical account of mathematics should satisfy. The semantic desideratum claims that "[a]ny theory of mathematical truth [should] be in conformity with a general theory of truth […] which certifies that the property of sentences that the account calls 'truth' is indeed truth" (Benacerraf 1973: 666). The epistemological desideratum requires that "[a] satisfactory account of mathematical truth […] fit into an over-all account of knowledge in a way that makes it intelligible how we have the mathematical knowledge that we have" (Benacerraf op. cit.: 667).

As is well known, those who abide by the semantic constraint take mathematical language at face value and are reluctant to wander away from what mathematicians literally say. In his 1965 article, Benacerraf did not picture the logicists as philosophers who wanted to satisfy the semantic constraint. Recall that the shared mistake of Ernie and Johnny was to refuse to take their teacher at his words and to replace the sequence of integers by a sequence of sets. In 1973, Benacerraf did not hold that logicism was motivated by "the concern for having a homogeneous

semantical theory in which the semantics for the statements of mathematics parallel the semantics for the rest of the language" (Benacerraf op. cit.: 661). Logicists did not hesitate to distort the surface form of arithmetical sentences.

On may rely on Shapiro (2006) for exactness so as to distinguish two readings of the semantic constraint. Let me quote Shapiro on this point:

> Before going further, I would like to acknowledge an orientation. As I see it, the goal of philosophy of mathematics is to interpret mathematics, and articulate its place in the overall intellectual enterprise. One desideratum is to have an interpretation that takes as much as possible of what mathematicians say about their subject as literally true, understood at or near face value. Call this the *faithfulness* constraint.
>
> Shapiro (2006: 110)

Shapiro then distinguishes "a second, and weaker, desideratum" from the faithfulness constraint:

> A second, and weaker, desideratum is to develop an interpretation that does not go too much beyond what mathematicians say about their subject. Surely the philosopher is going to say some things that the mathematician does not say. Mathematicians, as such, do not usually address philosophical issues about their subject. For example, they do not say much about what the natural numbers are, nor how we obtain mathematical knowledge, nor how mathematics applies to the physical world. Presumably, philosophical questions about mathematics are not to be answered solely in mathematical terms. The second desideratum is to not attribute *mathematical* properties to mathematical objects unless those attributions are explicit or at least implicit in mathematics itself. Call this the *minimalism* constraint.
>
> Shapiro loc. cit.

As an illustration of this last desideratum, Shapiro mentions Dedekind's abstractionist methodology:

> Richard Dedekind's philosophical methodology was to develop a system of objects, and then abstract the structure of the system. […] For example, he constructed the system of cuts in rationals, and then abstracted the real numbers from the cuts. According to Dedekind, the abstracted items — the real numbers — are not part of the system abstracted from, but are instead something "new" which the mind freely "creates." […] Dedekind's friend Heinrich Weber suggested instead that real numbers be *identified* with cuts. Dedekind replied that there are many properties that cuts have which would sound very odd if applied to the corresponding real numbers. […] For example, cuts have members. Do real numbers have members? Dedekind's Benacerraf-type point is consonant with the minimalism constraint.
>
> Shapiro op. cit.: 110–111

I take Shapiro's reference to Benacerraf to be a reference to Ernie's and Johnny's construals of the natural numbers: the two trainee logicists both make the Weber mistake—none of them respect the minimalism constraint.

What is, then, the correct reading of Benacerraf's 1973 semantic constraint? Should one treat it as strictly equivalent to the faithfulness constraint, or should one take it to consist in *both* the faithfulness *and* the minimalism constraints? In the former case, logicists would abide by the constraint since they construe numbers as objects (at least Frege would). But Benacerraf's (1965) whole point was to criticize

this weak reading. In 1965, Benacerraf was claiming that it is not acceptable to require that numbers be taken as objects since this kind of characterization might carry unwanted excess baggage, thereby making the choice between different definitions intractable. As Shapiro correctly points out, the bulk of Benacerraf (1965) was to say that logicists failed because they did not respect the minimalist constraint.

One should learn two lessons from this conclusion. First, even though Benacerraf remains silent on this point in 1973, one should ascribe to him a pretty strong reading of the semantic constraint. Recall that, in 1965, it was the minimalism constraint which carried the argumentative weight. The second lesson is that there was at the time, for Benacerraf, a tension between the logicist programme and the semantic constraint: logicists were charged indeed, in 1965, with violating minimalism. This second point is important because it shows that one cannot easily enroll logicism under the banner of philosophies that intend to satisfy the semantic constraint. But then, according to Benacerraf's 1973 dilemma, if a philosophy does not take the constraint seriously, it should at least try to abide by the epistemological constraint. Should one then interpret logicism as an epistemological program?

As we have seen in Sect. 6.1, this was precisely the view of logicism that the neo-positivists developed and that Benacerraf was opposed to in his 1981 article. Should we then ascribe to Benacerraf (1973) this neo-positivist interpretation of Frege's work?

In fact, neither Frege nor logicism play any role in 1973. When presenting an account of mathematical truth driven by epistemological considerations, Benacerraf refers to Hilbert and the combinatorialists—not to Frege and the logicists. True, there is one place where Benacerraf seems to say that what applies to combinatorialists applies to positivist logicists as well:

> Similarly, certain views of truth in arithmetic on which the Peano axioms are claimed to be "analytic" of the concept of number are also "combinatorial" in my sense. And so are conventionalist accounts, since what marks them as conventionalist is the contrast between them and the "realist" account [the one that respects the surface form of sentences].
>
> Benacerraf (1973: 665)

But Benacerraf does not mention Frege in this respect. And one obvious reason why Benacerraf might have to refer to Hilbert rather than Frege is that Hilbert's program was, unlike Frege's, explicitly put forward as an attempt to meet the epistemological challenge. It thus seems that in 1973 Benacerraf already had misgivings about the epistemologically oriented positivist interpretation of logicism. This would explain why he avoided then speaking about Frege and logicism.

In other words, my suggestion is that Benacerraf realized that logicism did not fit well with the dilemma he was trying to set forth. Logicism did not square with the philosophical frameworks which attempt to abide by the semantic constraint (this seems to underlie Benacerraf 1965's criticism), but neither can it be understood as having an epistemological program (this seems to underlie the criticism of Benacerraf 1981). This must be the reason why logicism disappeared from the scene. One may interpret the 1981 article as an attempt to make amends and fix one's previous mistake. If I am right, the impetus at the source of Benacerraf (1981)

comes from far away, namely from the difficulty to accommodate Frege to the framework set up in 1973.

## 6.3 Frege as a philosopher, Frege as a mathematician

Let me go back to Benacerraf (1981): the main claim is that, contrary to what neo-logicists think, Frege's logicist program was not driven by an epistemological concern. But then, what were the philosophical reasons that led Frege to espouse the logicist cause?

To answer that question, Benacerraf makes an important distinction between two kinds of foundationalist interests:

> A concern [with the foundations of arithmetic] might be interpreted in two different ways, corresponding to the interests of a philosopher and to those of a mathematician. Typically, the philosopher takes a body of knowledge as given and concerns himself with epistemological and metaphysical questions that arise in accounting for that body of knowledge, fitting it into a general account of knowledge and the world. That is Kant's stance. […] And that was the positivist's stance. But a mathematician's interest in what might be called "foundations" is importantly different. Qua mathematician, he is concerned with substantive questions about the truth of the propositions in question, as well as slightly more "philosophical" issues concerning how such propositions are properly established. The interests of the two groups are not disjoint — nor can these questions be sharply separated. But the differences are significant, and it is important to keep them in mind as we approach Frege.
>
> Benacerraf (1981: 23)

Indeed, according to Benacerraf, Frege came to adopt his logicism because he was moved by the mathematician's motivation:

> I claim that the Frege of the *Grundlagen* has the mathematician's motivation; that where he appears to deal directly with the more typically "philosophical" issues […], it is because he has restructured those questions and posed them in such a form that the answers they require will answer the substantive mathematical questions which are his principal concern.
>
> Benacerraf loc. cit.

The idea that one should distinguish between mathematical and philosophical motivations is something which is already present in 1965, when Benacerraf contrasts the philosophical demands of the logicists with the reasonable demands of the mathematicians:

> Martin [whom Benacerraf quotes at the beginning of the paper] correctly points out that the mathematician's interest stops at the level of structure. If one theory can be modelled in another […] then further questions about whether the individuals of one theory are really those of the second do not arise. In the same passage, Martin goes on to point out […] that the philosopher is not satisfied with this limited view of things. He wants to know more and does ask the questions in which the mathematician professes no interest. I agree. He does. And mistakenly so.
>
> Benacerraf (1965: 69)

What has changed in 1981 is that Frege, who was considered in 1965 as the paragon of a philosopher, is now regarded as a mathematician. In particular, in 1965, Benacerraf ascribed to Frege the wish to find the "real" essence of the mathematical object—that is, the wish to go beyond the level of structure. In 1981, Benacerraf claimed that, in Frege's *Grundlagen*, definitions are neither sense- nor reference-preserving. They are just like the usual mathematical definitions, whose sole aim is to pick out the entities (or the sets of entities) one wants to talk about, and which do not claim to deliver their true essence:

> I engaged in that discussion myself some years ago in a piece that can be read as arguing that either the definitions of the mathematical terms do not preserve their meaning, or their meaning does not determine their reference, since different and equally adequate definitions assign different referents to the mathematical vocabulary. I will argue later, contrary to what I formerly thought Frege to hold, that he and I speak with one voice.
>
> Benacerraf (1981: 18)

Frege is seen in Benacerraf (1981) not as someone who is interested in accounting for the nature of arithmetical knowledge, but as someone who seeks to extend mathematics by proving yet unproved elementary arithmetical theorems.

But this sharp distinction between the philosophical and the mathematical motivations is more slippery that it seems. Proving basic theorems of arithmetic requires holding that basic arithmetical truths could be grounded on yet more basic non arithmetical truths. In what sense exactly would they be more basic? Benacerraf makes it clear that the distinction between less and more basic has nothing to do with epistemology:

> The sense in which Frege [understands mathematics is one] that attempts to give some content to the notion of "the ultimate ground upon which rests the justification for holding […] a judgment to be true." For this is the metaphysical notion on which his view depends. I say "metaphysical" to contrast the dependence to which he is alluding with epistemic dependence. […] Frege is [concerned with] relations of dependence *among the propositions themselves*, whether or not they are believed and however those beliefs may be related to one another in the epistemic world of any individual. To prove a proposition involves (at least) deducing it from the propositions on which it "depends" in this metaphysical sense.
>
> Benacerraf op. cit.: 26–27

It is not easy to decipher what Benacerraf has in mind here, especially when the remark is relocated within his own framework, in which there is a sharp contrast between philosophically and mathematically oriented foundational programs. On the one hand, Benacerraf's new emphasis on the importance of finding new proofs seems to sever Frege from any metaphysical or philosophical concern. On the other hand, Benacerraf still acknowledges that Frege, when deriving arithmetic from logic, thought he had uncovered *the* true content of arithmetical propositions, as opposed to a mere particular method of proof. Frege's project is then not purely mathematical after all—as is shown by Benacerraf's reference to the notion of a "metaphysical" dependence between propositions. In other words, it seems that

Benacerraf still has trouble in 1981 characterizing the philosophical payoff of logicism in some unambiguous way.

There are some advances, however. Frege's logicism is not ignored as it was in 1973; it is indeed discussed. Benacerraf's stress on the importance of proofs moves Frege's logicism closer to Hilbert's combinatorialism. And the idea that Frege's basic propositions are not basic in an epistemological sense helps us to locate where exactly Hilbert and Frege (as characterized by Benacerraf) diverge: logicism is now sharply distinguished from epistemologically oriented philosophies of mathematics. Nevertheless, the emphasis being now on proofs rather than on sense- and reference-preserving definitions, Frege's logicism is thus disconnected from any philosophy keen on preserving the semantic content of mathematical statements. A close reading of Benacerraf (1981) thus confirms our previous exegetical hypothesis, namely that Frege's logicism, as read by Benacerraf, fails to meet both the epistemological challenge and the semantic challenge. From the point of view of the 1973 dilemma, it seems that Frege looses on both counts.

One might think that Benacerraf (1981) could provide us with the means to explain why this sad conclusion may not have bothered Frege too much. According to Benacerraf, Frege was not a philosopher of mathematics, but a mathematician. He didn't want to satisfy two incompatible desiderata; what interested him instead was to extend mathematics. To claim that Frege's project was first and foremost mathematical allowed Benacerraf to maintain his 1973 dilemma while finding a place for Frege's project which would be more in line with the historical Frege. But, as a result, this leads to restrict the scope of the 1973 dilemma, since, according to Benacerraf (1981), the dilemma does *not* concern Frege's philosophy of mathematics. Of course, this is so only because Frege's project is described by Benacerraf as being ultimately *not* philosophical. But this answer is very weak: according to the standard reading of Frege, Frege's work belongs indeed to the philosophy of mathematics. Benacerraf should then explain how to distinguish the "pure" philosophical project (facing the 1973 dilemma) from the hybrid mathematical-and-philosophical developments exemplified in Frege's works. Does the dilemma set forth in 1973 really exhaust the entire domain of the philosophy of mathematics? Or does it apply only to certain positions within a larger domain?

## 6.4 Benacerraf and Neo-logicism

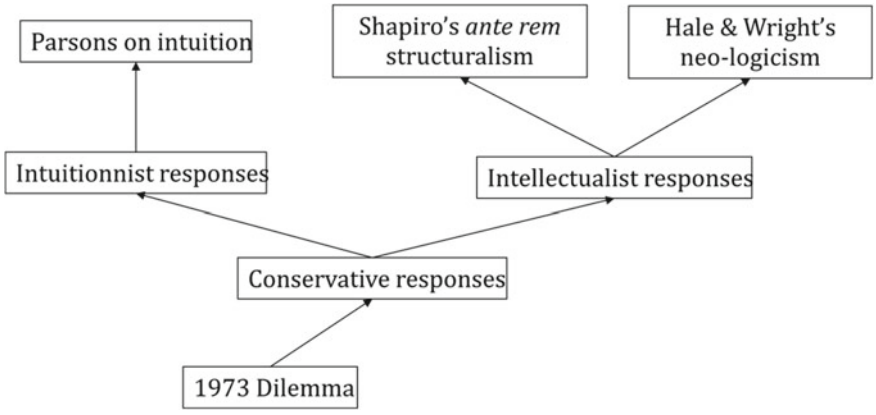
Before exploring the issue raised in the previous section, and in order to insist on the originality of Benacerraf's reading, I would like to contrast it with the neo-logicist interpretation of Frege. Benacerraf's (1973) dilemma is one of the starting points of the neo-logicists. Like the positivists, the neo-logicists regard Frege's work as constituting an epistemologically oriented program. But, unlike the positivists, they manage to recast Frege's construction so as to associate it with the

epistemological motivation. In this section, I will draw attention to a strange irony of fate: the dilemma that Benacerraf had trouble fitting into the Fregean framework is nothing but the dilemma that neo-logicists use as a starting point.

In their 2002 article, Hale and Wright classify the various kinds of responses to Benacerraf's challenge. They label conservative the responses "which take the Dilemma head-on, unrepently maintaining both that pure math is correctly construed at syntactic face value and that, so construed, it represents, at least for the greater part, a body of a priori knowledge" (Hale and Wright 2002: 103). They then distinguish two families of conservative responses: the intuitional and the intellectual. Let me quote from them:

> [Among conservative responses], two broad approaches seem possible: intuitional and intellectual. It may be proposed, first that an epistemology of mathematics should reckon with a special faculty — traditionally "intuition" — which enables an awareness of systems of abstracta […] broadly as ordinary sense perception makes us aware of ordinary concrete objects and their properties. Or it may be proposed that access to the objects of pure math is afforded by our general abilities of reason and understanding.
>
> Hale and Wright op. cit.: 104

Hale and Wright first establish that neo-logicism is an intellectual conservative response to the 1973 dilemma, and then attempt to show that this response fares better than its rivals—e.g. than Shapiro's *ante rem* structuralism. Here is the tree of the different possible conservative answers that Hale and Wright elaborate:




I won't be concerned here with Hale and Wright's arguments in favor of their own view. My point is simply that the neo-logicists conceive their project as an answer to Benacerraf's 1973 dilemma, i.e. as a way of meeting the epistemological challenge about the nature of arithmetical knowledge.

Despite the sharing of this conception with the neo-positivists, the neo-logicists' view is much more articulated than the positivists'. In particular, Hale and Wright do not adhere to the vague truth-in-virtue-of-meanings conception of analyticity.

The neo-logicist doctrine and its main difficulties have been presented and discussed in many papers; I shall restrict myself here to the main tenets of the program. What matters is that the neo-logicist's answer to the epistemological challenge raised by arithmetical knowledge is grounded in two key elements:

(1) The context principle, according to which "only in the context of a proposition does a word mean anything"—Hale and Wright use the principle as a device with which to settle worries about the reference to and knowledge of abstract objects[2]and
(2) Hume's Principle, which provides an explanation of how one has a grasp of the truth conditions of numerical identity statements, and hence of how one can acquire arithmetical knowledge.[3]

For the neo-logicists, the context principle and the abstraction schema (Hume's Principle) explain how we can refer to numbers and acquire arithmetical knowledge by reconceptualizing certain material equivalences without relying on intellectual intuition. Arithmetical statements are a priori in so far as they are deducible from second order logic extended by Hume's Principle. Arithmetical truths are thus not vaguely characterized as true in virtue of meaning. Neo-logicism is indeed a much more articulated conception than the neo-positivist view that Benacerraf attacked in 1981.

An interesting question then arises: if the neo-logicist interpretation of logicism is different from the neo-positivist one, should one still consider that there is an opposition between the former and Benacerraf's 1981 reading of Frege? Would Benacerraf approve the neo-logicist conservative response to his 1973 dilemma? Insofar as he insists on the fact that Frege's motivation was *not* epistemological, it seems that Benacerraf is opposed to the fundamental tenet of the neo-logicist interpretation of Frege. Although neo-logicism is much more refined than the neo-positivist doctrine, Benacerraf still claims in 1981, contra Hale and Wright, that Frege did *not* intend to account for mathematical knowledge, and if that is so, then neo-logicism doesn't quite provide a head-on response to the 1973 dilemma. This is ironical indeed since Hale and Wright root their own project in Benacerraf's dilemma. I shall devote the next section to a defence of this Benacerrafian interpretation (perhaps against Benacerraf's understanding of his own dilemma).

## 6.5 Semantic versus architectonic issues

We have seen that Benacerraf claimed in 1973 that philosophy of mathematics falls prey to a dilemma, but that, in 1981, he seemed to take half a step back with the implication that Frege developed a hybrid mathematical and philosophical project. I also pointed out that Benacerraf's 1981 characterization of Frege's project

remains unstable: Frege's research was then seen as a purely mathematical one while it was still grounded in purely metaphysical considerations. I will suggest now that this last difficulty comes from the fact that Benacerraf did not go far enough in his criticism of his 1973 dilemma. What Benacerraf says about Frege in 1981 should have led him to cast doubt on one of its horns, namely the semantic horn.

The argument to that effect is as a matter of fact quite simple. According to Benacerraf, Frege's analysis of arithmetic does not take arithmetic at face value, at least if the notion of face value is to be taken in Shapiro's substantial sense (see Sect. 6.2). But according to Benacerraf (1981), Frege's heterodox analysis of arithmetical statements was a central part of the new mathematics he was busy developing. So in reshaping arithmetical statements as he did, Frege was faithful indeed to a certain kind of mathematics after all, namely to the logical calculus he was elaborating. Did Frege abide by the semantic constraint of the dilemma, or did he not? Even if he wasn't faithful to ordinary arithmetic, Frege was certainly faithful to the new mathematics he created.

This example shows that the target that philosophers should aim at is much less definite than Benacerraf seems to believe. There is nothing trite in the task of giving an unambiguous characterization of the standard form of mathematical discourse. The Fregean mathematician, against his arithmetician colleagues, will consider that "There are at least three perfect numbers greater than 17" does *not* have the form "There are at least three *FG*'s that bear *R* to *a*." So which mathematician should a philosopher follow? This sort of disagreement about the best way to formulate a theorem is not an exception among mathematicians. What is the genuine form of the fundamental theorem of arithmetic? Does the result merely say something about integers, or does it implicitly allude to Riemann's function zeta (since a reference to it seems needed in its proof)?[4] What is the genuine form of the fundamental theorem of algebra? Does it say something about the intersection of algebraic curves and nothing more, or does it implicitly refer to topological considerations?[5] What is the genuine form of the fundamental theorem of real projective geometry? Does it say something about projective order, as Klein believed, or does it merely say something about the incidence structure of the real projective plane?[6] In all of these cases, mathematicians disagree on the formulation of some key results, and these disagreements, far from being merely rhetorical, encapsulate important differences in the way one views mathematics as a whole, and in the way one anticipates its future progress.

Disagreements and conflicts about the correct way to analyze a mathematical proposition play an important role in mathematics. Benacerraf's idea that Frege's logicization of arithmetic should be seen, first and foremost, as a mathematical project is in this respect quite legitimate. By reconceptualizing basic arithmetical statements, Frege imitates what mathematicians never stop doing in other areas. But this diagnosis should have led Benacerraf to cast doubt on the legitimacy of the semantic horn of his dilemma. If an integral part of mathematical activity consists in reconceptualizing mathematical discourse, then it is no longer possible to believe that mathematicians adhere to a mathematical language that has been fixed once and

for all. It thus becomes difficult to describe the semantic horn of the dilemma in terms of a faithfulness demand in Shapiro's sense. To be faithful to mathematics, philosophers should rather account for the fact that there are some cases where mathematicians do *not* agree on the way a theorem should be stated.

Mark Wilson has forcefully defended this point. Wilson remarks that the linguistic maturity of a science often arises "after its basic terminology seems to have settled into fixed meanings" (Wilson 1994: 520). Wilson's point is to explain that in such cases the apparent grammar of a proposition is gradually superseded by a (at first) hidden grammar, which gradually emerges from the original language after repeated experiences. Wilson calls this second grammatical pattern the "working grammar":

> [In the process] agents discover that inferential pathways validated in their original apparent semantics sometimes lead to unhappy results, whereas other, officially unsanctioned, deductions generally lead to success. At first, the relevant speakers tend to excuse these deviations from apparent semantical correctness by a variety of ad hoc explanations. Over time, however, a *system* seems to emerge within the deviations and the agents now recognize, as they reassess their language's workings, that a distinct grammatical parsing better captures the inferential successes and failures they have encountered. Previously unrecognized syntactic categories now become salient as the keys to deductive validity. Call this second pattern of grammatical structuring, which slowly emerges from the language after repeated experience, a *working grammar* for the language. The reappraisal of word/world relations which belongs with this new grammar becomes the language's *working semantics*.
>
> Wilson loc. cit.

The disagreements among mathematicians we previously alluded to precisely concern the ways in which one could devise a working grammar for a given mathematical language. In fact, the very distinction between an apparent and a working grammar is a challenge for Benacerraf's semantic challenge: which grammar are we supposed to follow as far as philosophy is concerned, the first one or the second one?

This point also pertains to Benacerraf's global conception of the relation between philosophers and mathematicians. Benacerraf (and this is something which remain stable from 1965 to 1981) describes life within the mathematical community as peaceful and consensual. According to him, mathematicians are always in agreement about the scientific issues. The troublemakers are outsiders—philosophers, who are motivated by the epistemological goal consisting in locating mathematical knowledge within knowledge in general and whose business is not scientific progress. According to Benacerraf, philosophers introduce unwarranted divisions when asking mathematicians irrelevant questions. In this regard, Benacerraf is very close to the neo-positivists he criticized in other respects—for him as for them science is metaphysically neutral, and it is the job of the "good" philosophers to protect scientists against the pernicious influence of the metaphysicians. It is in this particular historical context that Benacerraf's emphasis on the demand that mathematical discourse be taken at face value should be understood. The semantic constraint plays the role of a safety net, which prevents philosophers from falling into metaphysics.

It seems to me that Benacerraf somehow exaggerates the unity of the mathematical community. Of course, mathematicians do agree on the fact that a certain proof proves a certain theorem.[7] But they often disagree on the best way to analyze mathematical propositions, and on their particular place and role within the architecture of the mathematical sciences. The semantic constraint is indifferent to the fact that there is no consensus, within the mathematical community, about the way one should view the global organization of mathematics. Grammar is not a reliable guide to answer this question. Algebraists do not compare mathematical theories in the same way as analysts or geometers. These disagreements are not caused by the malign influence of intruders—they come from the development of mathematics itself. How then to settle such internal disagreements between mathematicians? The answer is: by doing mathematics, i.e. by reshaping the standard theorems and the standard proofs in a way that conforms to our preferred view, but also by elaborating philosophical justifications about the relations between the various mathematical disciplines.

In brief, what I want to suggest here is that, beside the epistemological problem (how to locate mathematical knowledge within the totality human knowledge?), there is another issue I shall call the architectonic issue, which deals with the articulation of the various mathematical disciplines and the formulation of their main theorems. It seems that one cannot approach these two questions simultaneously. The wish to locate mathematical knowledge within human knowledge in general leads to homogenize mathematics, and thus to lose sight of the internal diversity of mathematics. Conversely, the attempt to compare various ways of organizing mathematics tends to hide what makes mathematical knowledge, taken as a species of knowledge, so specific (the fact that it deals with abstract objects, for instance). In this respect, the semantic constraint acts as a powerful homogenization operator. As a consequence, when abided by too strictly, the constraint forbids articulating the architectonic issue. The semantic requirement makes us forget that, in mathematics, there are more than one way to articulate a theorem, and that such differences are mathematically relevant.

Once the existence of the architectonic issue is acknowledged, an explanation of Frege's talk about the "metaphysical" dependence between propositions may be provided. Frege strives to promote a Gaussian vision of mathematics in which arithmetic (the queen of the sciences) is neatly distinguished from the other mathematical disciplines, especially geometry. The wish to guarantee a special place for arithmetic within mathematics could be the impetus behind the idea that arithmetical theorems, unlike geometrical propositions, depend on the most general laws of thought. The idea that certain propositions depend (in a non epistemological sense) from others might be understood as a way of describing the relation between the different disciplines within mathematics.

I won't defend this exegetical claim here. It is worth being mentioned because it seems to prolong Benacerraf's 1981 interpretation. It does share with Benacerraf's reading the idea that Frege's project was not inherently epistemological. This is an important point of agreement because it distinguishes these readings from both the neo-positivist and the neo-logicist readings. But the architectonic interpretation differs

from Benacerraf's in that it refuses to oppose mathematical and philosophical issues; some questions (the architectonic ones) are inextricably mathematical and philosophical. Contrary to what Benacerraf (1981) implies, it isn't sufficient to show that a project isn't epistemological to show that it is purely mathematical. As a matter of fact, Benacerraf's faithfulness to the semantic constraint prevents him from recognizing the existence of the architectonic interpretation, i.e. of the convergence and interdependence of the mathematical and the philosophical. In still other words, I think that Benacerraf recognizes in 1981 that Frege, in transforming the grammar of usual arithmetic, acted as a mathematician. It nevertheless seems to me that Benacerraf doesn't draw the full lesson he should have drawn. His analysis should have led him to challenge his own faithfulness to the semantic constraint, just as it should have led him to recognize the existence of hybrid—i.e. both mathematical and philosophical—questions. It is precisely because he refrained from drawing these conclusions that he failed to identify the architectural motivation behind Frege's logicism.

## 6.6 Conclusion

I have been making two claims in this paper. The first claim is that Benacerraf (1981) should be viewed as the conclusion of a move, which sprung from Benacerraf's early criticism of Frege in 1965, and passed through the elaboration of his dilemma in 1973. Although Benacerraf did not identify Frege's program as epistemological, he considered that Frege did not abide by the semantic constraint. Logicism was then for him a puzzle. The second claim is more general: in order to be consistent, Benacerraf should have put his semantic constraint into question in 1981. Benacerraf acknowledges that the two horns of the dilemma do not exhaust the possible philosophical positions, but his faithfulness to the dilemma leads him to expel Frege's work from the sphere of the philosophy of mathematics. I think on the contrary that Frege's undertaking could be read as an attempt to solve the architectonic issue (how should we view the organization of mathematics?) and that this question is a philosophical as well as a mathematical one.

Concerning this second claim, let it be clear that I don't mean to undermine the importance of Benacerraf's (1973) dilemma. Mathematics does indeed raise an epistemological issue: mathematical knowledge exhibits features that make it utterly different from other kinds of knowledge, and this, in and of itself, raises a whole set of intricate puzzles. Neo-logicism, *ante rem* structuralism, etc. strive to provide answers to a venerable family of deep problems. My worry does not concern the epistemological constraint as such, but only a reading of the semantic constraint that is way too strong.[8] Taken too seriously, the constraint can lead us to believe that the only reason we have to change the surface form of mathematical sentences comes from the wish to find a reasonable epistemology for mathematics. This simply isn't true: mathematicians never stop modifying their notation and reconceptualizing the contents they express. As Hale and Wright have noted, however, the epistemological issue persists even if we do not strictly follow the

surface form of mathematical statements. So after all, perhaps we don't need the semantic constraint—or perhaps a very weak version of it will do. But it seems to me that the quite radical way in which Benacerraf has formulated his constraint in 1973 has contributed to stray contemporary philosophical attention away from mathematico-philosophical issues about the internal organization of mathematics. And I think this is a pity.

**Notes**

1. *See* Benacerraf (1981: 24): "The point of [*Grundlagen*, §3] is to separate the notion of the content of a judgment from that of the justification for the judgment—in the sense of justification introduced in [*Grundlagen*, §2]; namely, the "support" of the judgment; the propositions on which it 'depends' for its truth."
2. "[According to the context principle,] singular thought, and object-directed thought in general is […] *enabled and fully realized in* an understanding of suitable kinds of statements. […] The opposite idea is precisely what is embodied in the Augustinian conception of language put up for rebuttal at the very outset of the *Philosophical Investigations*; and the prime spur towards the 'naturalist' tendency which finds abstract objects per se problematical is the idea, at the heart of the Augustinian conception, that some, however primitive, form of conscious acquaintance […] must lie at the roots of all intelligible thought of, and hence reference to objects of a particular kind" (Hale and Wright 2002: 115–116).
3. "The *import* of the stipulation of the equivalence is simply that corresponding instances of the left and right sides—matching sentences of the shapes 'the direction of line a = the direction of line b' and 'lines a and b are parallel'—are to be alike in truth-value, i.e. materially equivalent. But because the stipulation is put forward as an explanation, its *effect* is to confer upon statements of direction-identity the *same truth conditions* as those of corresponding statements of line-parallelism. Thus what a recipient of the explanation immediately learns is that whatever suffices for the truth of a statement of line-parallelism is equally sufficient for the truth of the corresponding statement of direction—identity. However, she also understands that she is to take the surface syntax of direction-identity statements at face value. She already possesses the general concept of identity, and so is able to recognize that the expressions flanking the identity sign must be singular terms. […] [From this,] she learns […] that directions just are objects with exactly those identity-conditions, and thus acquires the concept of *direction*" (Hale and Wright 2002: 117–118).
4. For a discussion of the fundamental theorem of arithmetic, see Arana (2011).
5. For an overview of the various proofs of the theorems and the many ways of looking at complex polynomials, see Fine and Rosenberger (1997).
6. See Gandon (2012): Chaps. 1 and 2.
7. One might wish to challenge this by referring to intuitionist mathematics, or to issues raised by computer-assisted proofs.
8. This is the kind of reading that Shapiro favors in Shapiro (2006). See Sect. 6.2.

## References

Arana, A. (2011). L'infinité des nombres premiers : une étude de cas de la pureté des méthodes. *Les études philosophiques,2*(97), pp.193–213.

Benacerraf, P. (1965) What Numbers Could Not Be, *The Philosophical Review, 74*, 47–73.

Benacerraf, P. (1973) Mathematical Truth, *The Journal of Philosophy, 70*, 661–679.

Benacerraf, P. (1981) Frege: The Last Logicist, *The Foundations of Analytic Philosophy,* Midwest Studies in Philosophy, *6*, l7–35.

Fine, B., & Rosenberger, G. (1997). *The fundamental theorem of algebra*. New York: Springer.

Gandon, S. (2012). *Russell's unknown logicism*. Basingstoke: Palgrave MacMillan.

Hale, B., & Wright, C. (2002). Benacerraf's dilemma revisited. *European Journal of Philosophy, 10*(1), 101–129.

Musgrave, A. (1977). Logicism revisited. *The British Journal for the Philosophy of Science, 28*(2), 99–127.

Quine, W. V. O. [1935]. (1976). Truth by convention. In *The ways of paradox and other essays* (Revised and enlarged edition, pp. 77–106). Cambridge, MA: Harvard UP.

Shapiro, S. (2006). Structure and identity. In F. MacBride (Ed.), *Chapter 5 of Identity and modality. New essays in metaphysics* (pp. 109–145). Oxford: Oxford UP.

Wilson, M. (1994). Can we trust logical form? *The Journal of Philosophy, 91*(10), 519–544.

# Chapter 7
# Truth, Fiction, and Stipulation

**Mary Leng**

For the past few years, I have been fortunate enough to teach, annually, a third year undergraduate module in the philosophy of mathematics. It is a testimony to Paul Benacerraf's great influence on the discipline that the module is structured very naturally in two halves, which could quite easily be subtitled "Before Benacerraf" (BB) and "After Benacerraf" (AB). The story I tell starts at the end of the 19th century, with Cantor's development of the new infinitary set theory, and mathematicians' and philosophers' concerns about how (or whether) we can make sense of this new mathematics that is not grounded in Kantian intuition of space and time. We look at the usual "big three" of logicism, intuitionism, and formalism, and consider how, during this period, it was possible, and perhaps even quite plausible, to think (as Hilbert did in his debate with Frege over the nature of axioms) of mathematical truth as something quite different from empirical truth, or truth *simpliciter*. Truth in mathematics could plausibly be thought of as an internal affair, a matter of mere derivability from consistent axioms, rather than correspondence to some external mathematical reality.

Along, alas, came Kurt Gödel, whose first incompleteness theorem put paid to the simple equation of truth with formal derivability found in Hilbert's early discussions about mathematics, by presenting a statement of arithmetic such that neither it nor its negation was formally derivable (and that, nevertheless, we have good reasons for taking to be true). Furthermore, Gödel's second theorem destroyed the hopes for Hilbert's more mature view of mathematics according to which finitary mathematics has a meaningful and true interpretation (in terms of finite strings of strokes), but infinitary mathematics should be viewed as a strictly meaningless "useful instrument," consisting of ideal elements introduced to smooth out our theory of the finitary meaningful portion of mathematics. Hilbert's "one condition, albeit an absolutely necessary one" ([1926] 1983: 199) on this account

M. Leng (✉)
Department of Philosophy, University of York, Heslington, YO10 5DD, UK
e-mail: mary.leng@york.ac.uk



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_7

was the provision of a finitary (and therefore meaningful) consistency proof for infinitary "ideal" mathematics, to show that this "instrument" would not conflict with true infinitary claims. Gödel's second incompleteness theorem showed that the consistency proof that Hilbert envisaged was impossible.

One upshot of Gödel's 1931 paper, so I tell my students, is thus that it invited mathematicians and philosophers to take seriously mathematical truth as a species of genuine truth, i.e., not as reducible to formal derivability. And this they did, with things going rather quiet on the formalist front post-Gödel. Gödel's own platonism, presented, e.g., in his 1947 paper "What is Cantor's Continuum Problem?" (Gödel [1947] 1990), sees mathematics as a science alongside the physical sciences, with similar methods of justification available (such as use of the hypothetico-deductive method to justify axiom choices in light of their consequences, as well as, in what is now most often quoted with an accompanying eye-roll, Gödel's claim that "despite their remoteness from sense experience, we do have something like a perception also of the objects of set theory, as is seen from the fact that the axioms force themselves on us as being true" (Gödel [1964] 1990: [271] 268)). Mathematical truth thus becomes answerable to mathematical reality in the same way that empirical truths are answerable to physical reality. And with this picture in place, it was only a matter of time before someone would put 2 and 2 together and wonder how we can be said to know those things we take ourselves to know about this realm of mathematical objects.

While 1931 brought the metaphorical death—via Gödel's results—of Hilbert's programme, it also brought the very real birth, in Paris, of Paul Benacerraf. Events in Europe meant that Benacerraf and Gödel both found their ways to the US, with Gödel leaving Austria to join the Institute of Advanced Study at Princeton in 1940, where he remained until his retirement in 1976. Paul Benacerraf was still a child when he left Paris for the US, but it was not long before he too found his way to Princeton, as an undergraduate (graduating with his A. B. in 1953), then Ph.D. student and, from 1960, a Faculty member in the Department of Philosophy. The story I tell in my lecture course imagines the young Benacerraf hearing of the old Gödel's platonism, digesting this as the new orthodoxy, puzzling about it, then tapping him on the shoulder around 1965 (with "What Numbers Could Not Be") and again around 1973 (with "Mathematical Truth") to say, "But isn't that view just crazy?", thus setting in place the predicament for philosophy of mathematics ever since. On the one hand, the failure of the "less substantial" views of mathematical truth suggests that platonism has to be right, but on the other hand, the sheer craziness of platonism (and in particular the epistemic puzzle Benacerraf raises for platonism) suggests that it cannot be so. Philosophy of mathematics AB has largely consisted of attempts to pick up the pieces in light of this attack on the platonist orthodoxy and the apparent lack of any acceptable alternative.

Such is my story, but it is of course just a story—at best a rational reconstruction of some developments in the history of the philosophy of mathematics through the 20th century. It makes for a nice narrative, and a neat way of organizing a lecture course, but each year as I present it I become less happy with the simplistic picture that arises. In particular, I worry, by centering on events around Princeton it ignores

what's going on in Harvard in the BB period, with the work of W.V.O. Quine. Of course, I do cover Quine's work, fitting it in with more recent discussions of the indispensability argument AB. There, Quine's indispensability considerations are presented as yet another reason to take platonism seriously, but Quine's holistic epistemology, I claim, does not have the resources by itself to solve the epistemic puzzle raised by Benacerraf in "Mathematical Truth," at least not if this is recast, as Field (1989) does, in terms of the puzzle of explaining the reliability of our mathematical beliefs. The puzzle remains, I claim—indeed, it is made even more stark, because indispensability considerations give us even more reason for taking mathematical truth seriously, but provide no account of how it could be that beliefs gained through our attempts to organize our empirical experiences could manage to get things right about a causally isolated realm of mathematical objects.

But—and now for confessions—I find myself increasingly unhappy with this account of the position of Quine's views in relation to Benacerraf's 1973 puzzle. I fear that its initial plausibility depends on a metaphysically heavyweight understanding of "objects," coming out of Gödelian platonism, that Quine himself would be unwilling to accept. Benacerraf's 1973 paper presents a dilemma for any account of mathematical truth. On the one hand, adopt a "standard" platonist semantics that interprets mathematical truths as truths about a realm of abstract objects, and face a difficulty explaining how we could come to know any mathematical truths; or, on the other hand, adopt an alternative "combinatorial" semantics that accounts for mathematical truths in terms of our methods of coming to know them (through proof from stipulated axioms), and face a difficulty explaining how this could possibly be enough to bring genuine truth. Understood as an attempt to provide an epistemology for platonism as construed in the "standard" view (which sees mathematics as consisting in a body of truths about an independent realm of mathematical objects), then Quine's holism does seem to fall short of answering the puzzle Benacerraf raises and Field revives ("how could beliefs gathered in *those* ways get things right about objects like *that*?"). But, reading Quine carefully, it is perhaps better to understand him as answering (before the fact!) Benacerraf's challenge not to the standard view, but to combinatorial views, of explaining how postulational stipulation can—if the circumstances are right, and despite Benacerraf's protestations to the contrary—"provide for truth" (Benacerraf 1973: 679). So my worry is that it begins to look as though Quine's philosophy of mathematics provides the resources to answer Benacerraf's dilemma, and did so even before Benacerraf presented his challenge in his 1973 paper. Does this mean that the philosophy of mathematics AB has been sent down a wrong path, expending endless efforts on solving a problem that was already solved BB, if only we'd paid attention to Harvard rather than Princeton, with a careful reading of Quine?

Well, I don't think so (though, as a philosopher of mathematics fully entrenched in the AB tradition, I would say that, wouldn't I?). But I have come to think that Quine's view requires a response in its own terms, read as a challenge to Benacerraf's claim that "stipulation does not provide for truth", rather than as an attempt to fill in an epistemology for standard platonism that bridges the gap between acausal abstracta and causally located human knowers. So what follows is my

attempt at telling a new story about the Quinean "solution" to the problems raised in "Mathematical Truth," one that, I hope, I'll be able to tell my students in the future without wincing.

## 7.1 Benacerraf's Problem for Combinatorial Views

Let us go back, then, to Benacerraf (1973), and Benacerraf's challenge, not to the Gödelian platonist, but to those who wish to ground mathematical truth in derivability from stipulated axioms, providing what Benacerraf calls a "combinatorial" view of mathematical truth. "Mathematical Truth" may be best remembered for its epistemic objection to the standard (platonist) view of mathematical truth, but it is with his objection to combinatorial views that Benacerraf chooses to end his paper. Benacerraf is, of course, aware of the worries that the first incompleteness theorem raises for taking truth in mathematics to be grounded in derivability from stipulated axioms, but he does not take these to be conclusive. Neither is he convinced by the completeness of Quine's objections to views of this sort in "Truth by Convention" (Quine [1935] 1976), where it is argued that it cannot be conventions "all the way down," since this objection applies to the thesis of the conventionality of logic, without getting to the heart of the problem with taking mathematical truth as grounded in logic plus conventions. Neither Gödelian worries nor Quine's objections get to the heart of what, in Benacerraf's eyes, is wrong with the idea of "truth by convention," since both would allow for the conventional stipulation of some truths. "The deeper reason," Benacerraf tells us, "is that postulational stipulation makes no connection between the propositions and their subject matter—stipulation does not provide for truth" (Benacerraf 1973: 679). The paper ends:

> To clarify the point, consider Russell's oft-cited dictum: "The method of 'postulating' what we want has many advantages; they are the same as the advantages of theft over honest toil" (Russell 1919: 71). On the view I am advancing, that's false. For with theft at least you come away with the loot, whereas implicit definition, conventional postulation, and their cousins are incapable of bringing truth. They are not only morally but practically deficient as well.
>
> Benacerraf loc. cit.

## 7.2 The Fictionalist's Stance

The "method of 'postulating'" that Benacerraf complains about has its contemporary expression in mathematical fictionalism. In relation to unapplied mathematics, fictionalists see mathematicians as involved in working out the logical consequences of stipulated axioms. And insofar as fictionalists see mathematical correctness as a matter of following logically from such axioms, the view is

"combinatorial" in character (albeit generally involving a beefed up modal notion of "logical consequence," rather than resting mathematical correctness in formal derivability). But Benacerraf's complaint against "the method of 'postulating'" is well taken by fictionalists, who do not claim to "come away with the loot" of mathematical truth. Fictionalists agree that one cannot make axioms true by fiat. So insofar as mathematical theories are grounded in stipulated axioms, we have no reason to call such theories true.

Indeed, Quine too agrees that this understanding of mathematical theories does not bring truth, in a passage that predates Benacerraf's own complaint that "stipulation does not provide for truth":

> Playing within a non-Euclidean geometry, one might conveniently make believe that its theorems were interpreted and true; but even such conventional make-believe is not truth by convention. For it is not really truth at all; and what is conventionally pretended is that the theorems are true by non-convention.
>
> Quine [1954] (1977: 116)

Yet, despite all he has to say against the concept of truth by convention, there is a sense in which Quine can be viewed as arguing, against Benacerraf, that if the circumstances are right then stipulation can and does bring truth with it.

## 7.3 Truth by Convention from the Enemy of Truth by Convention?

When, for Quine, does stipulation lead to truth? All the time, in fact! One way into Quine's rejection of the analytic/synthetic distinction is via the recognition of the presence of conventional/stipulated elements in all truths ("Taken collectively, science has its dual dependence on language and experience; but this duality is not significantly traceable into the statements of science taken one by one" (Quine [1951] 1953: 42)). It may be a feature of some sentences that they were introduced entirely as a matter of conventional stipulation. But this, in Quine's view, is not a lasting feature, or a particularly important one. "Conventionality," Quine tells us,

> is a passing trait, significant at the moving front of science but useless in classifying the sentences behind the lines. It is a trait of events and not of sentences. […] Legislative postulation contributes truths which become integral to the corpus of truths; the artificiality of their origin does not linger as a localized quality, but suffuses the corpus.
>
> Quine [1954] (1977: 119–120)

The development of science is, in Quine's view, a matter of stipulation leading to truth, again and again, as convenient ways of speaking prove themselves to be useful in organizing experience, and through that receive empirical confirmation as part of a theoretical package.

But don't the axioms of mathematical theories have a quite different character to conventions adopted in empirical science? Einstein presents the assumption that

light travels at the same speed in all directions as "in reality neither a supposition nor a hypothesis about the physical nature of light, but a stipulation which I can make of my own freewill in order to arrive at a definition of simultaneity" (Einstein 1920: 23). But although we cannot arrange a direct test of the one-way speed of light (the best we can do is measure speed over a round-trip), one might think that the utility of the theoretical package as a whole that starts from the stipulation that the speed of light is the same in all directions is enough for us to take the package as a whole, including its conventions, as confirmed as true (if anything is), despite its starting point in an explicit stipulation. The axioms of set theory might look different in this respect: they are adopted apparently without any view to empirical matters, and appear immune from empirical testing in a quite different way. We can make sense of the idea of testing the speed of one-way light signals, using synchronized clocks at a distance, but our difficulty is that we cannot ensure clocks are synchronized without making assumptions about the one-way speed of light, so from a practical perspective we are stuck with adopting an assumption as a convention. But it's simply unclear what could count as an empirical test of the axioms of set theory, even in ideal conditions, since they do not appear to be making any claims about empirical matters.

This Quine disagrees with fundamentally, holding firstly that the conventional character of the axioms of set theory is just the same as the conventionality found elsewhere, and secondly, that the mathematical components of our theories are ultimately in receipt of empirical confirmation through their presence in an entire theoretical package.

> What seemed to smack of convention in set theory […], at any rate, was "deliberate choice, set forth unaccompanied by any attempt at justification other than in terms of elegance and convenience"; and to what theoretical hypothesis of natural science might not this same character be attributed? For surely the justification of any theoretical hypothesis can, at the time of hypothesis, consist in no more than the elegance or convenience which the hypothesis brings to the containing body of laws and data. How then are we to delimit the category of legislative postulation, short of including under it every new act of scientific hypothesis?
>
> The situation may seem to be saved, for ordinary hypotheses in natural science, by there being some indirect but eventual confrontation with empirical data. However, this confrontation can be remote; and, conversely, some such remote confrontation with experience may be claimed even for pure mathematics and elementary logic. The semblance of a difference in this respect is largely due to overemphasis on departmental boundaries. For a self-contained theory which we can check with experience includes, in point of fact, not only its various theoretical hypotheses of so-called natural science but also such portions of logic and mathematics as it makes use of. Hence I do not see how a line is to be drawn between hypotheses which confer truth by convention and hypotheses which do not, short of reckoning all hypotheses to the former category.
>
> Quine [1954] (1977: 121–122)

In science, then, stipulation leads to truth all the time. What starts as stipulation is confirmed as true through the success of theories grounded in that stipulation in empirical theorizing.

In light of Quine's blurring of "departmental boundaries," and his placing of mathematics alongside the rest of empirical science when it comes to confirmation, Benacerraf's epistemological objection to the "standard view" of mathematical truth seems to miss the mark. Benacerraf complains that "the principal defect of the standard account is that it appears to violate the requirement that our account of mathematical truth be susceptible to integration into our over-all account of knowledge" (Benacerraf 1973: 670).

But Quine's account of the confirmation of empirical theories is designed precisely so as to integrate our knowledge of mathematical truths into our over-all account of knowledge. Benacerraf continues:

> If, for example, numbers are the kinds of entities they are normally taken to be, then the connection between the truth conditions for the statements of number theory and any relevant events connected with the people who are supposed to have mathematical knowledge cannot be made out.
>
> Benacerraf (1973: 673)

But in Quine's view, the idea that our knowledge of objects requires a connection between the knower and the objects known is just mistaken. Physical objects, as well as mathematical, are introduced as convenient ways of organizing experience, and our beliefs about them are confirmed by the serviceability of that convention:

> Physical objects are conceptually imported into the situation as convenient intermediaries not by definition in terms of experience, but simply as irreducible posits comparable, epistemologically, to the gods of Homer. For my part I do, qua lay physicist, believe in physical objects and not in Homer's gods; and I consider it a scientific error to believe otherwise. But in point of epistemological footing the physical objects and the gods differ only in degree and not in kind. Both sorts of entities enter our conception only as cultural posits. The myth of physical objects is epistemologically superior to most in that it has proved more efficacious than other myths as a device for working a manageable structure into the flux of experience.
>
> Quine [1951] (1953: 44)

Quine's philosophical outlook, developed in his writings in the 1950s, thus appears to provide the unified account of the semantics and epistemology of mathematics that Benacerraf, in 1973, claimed was lacking. By providing a route from conventional stipulation to truth (via successful use of stipulated assumption in empirical science), Quine was able to argue that, despite their roots in explicit convention, mathematical hypotheses could be as well-confirmed as any of the hypotheses of our empirical scientific theories, and that our knowledge of mathematics was of a kind with our knowledge of physical objects. Both are stipulated in the course of our attempts to organize our experience, and both are vindicated by the success of such attempts.

## 7.4 What's Wrong with Quine's "Solution"?

It looks, then, as though Quine had the solution to Benacerraf's famous problem well before Benacerraf wrote "Mathematical Truth." Why, then, have Benacerraf's concerns about the difficulty of reconciling one's account of mathematical truth with an epistemology for mathematics been so hugely influential? One reason is that they were developed without Quine's holistic epistemology in mind, in the heyday of the causal theory of knowledge. Against the target of epistemological accounts along those lines (i.e., that ask what extra ingredients may be needed to make an individual justified, true, belief count as knowledge), Benacerraf's concerns surely hit their mark. But even when alternative epistemologies are brought into view, Benacerraf's knowledge problem still seems to pack a punch. One influential way of recasting Benacerraf's problem in the absence of a causal theory of knowledge is Hartry Field's:

> The way to understand Benacerraf's challenge, I think, is not as a challenge to our ability to justify our mathematical beliefs, but as a challenge to explain the reliability of these beliefs […]. Benacerraf's challenge — or at least, the challenge which his paper suggests to me — is to provide an account of the mechanisms that explain how our beliefs about these remote entities can so well reflect the facts about them.
>
> Field (1989: 25)

And, as I have said, it is this response that I quote to my students when trying to explain why Quine's indispensability argument does not provide a satisfactory solution to Benacerraf's knowledge problem. Even if the indispensability considerations tell us that we ought to believe in mathematical objects, they still leave it entirely mysterious as to how beliefs reached in *that* way (through our attempts to organize our experience of the physical world) should get things right about objects of *that* sort (acausal, nonspatiotemporal, mind- and language-independent abstracta).

My concern with this quick response to Quine is that, in its worries about "remote entities," it takes objecthood too seriously, and does not take seriously enough the conventionality of all "objects" in Quine's theoretical framework. In Quine's view, *all there is for there to be evidence for the existence of objects* is for our object-involving conventions to be serviceable ones. A "reliability" explanation, linking our evidence for believing in $\phi$s to the facts about $\phi$s, is not needed to underpin our knowledge of mathematical objects, simply because such an explanation is not needed anywhere. We have reason to believe in electrons because the electron hypothesis is serviceable, and the same goes for numbers or tables and chairs. By focussing on the "remoteness" of mathematical objects, Field's characterisation of the knowledge problem simply ignores Quine's rather thin conception of objecthood.

## 7.5 Why I Am not a Quinean: Confirmation, Explanation, and Good Old Causal Isolation

Quine's work throws into doubt the Benacerrafian concern that we could never have knowledge of abstract objects because of the lack of a causal connection between us and the objects known. In Quine's view, knowledge of objects of a particular kind doesn't require any such causal connection; it just requires that our beliefs about such objects form part of a well confirmed scientific theory. In Quine's holistic view, whatever conventional stipulations remain in our best scientific theory should be considered as confirmed truth, since confirmation extends to the entire theoretical package, conventions and all.

However, as has been emphasized in recent challenges to the indispensability argument (Maddy 1992, 1997 and Leng 2010), there appear to be plenty of cases where we successfully adopt a convention to speak as if there were $\phi$s without having reason to think that the existence of $\phi$s is thereby confirmed. In fluid dynamics, for example, we choose to speak as if fluids were continuous substances, even though we know they are no such thing. Indeed, in his textbook presentation of the theory, G. K. Batchelor stipulates that he will adopt the following "continuum hypothesis":

> […] that the macroscopic behaviour of fluids is the same as if they were perfectly continuous in structure; and physical quantities such as the mass and momentum associated with the matter contained within a given small volume will be regarded as being spread uniformly over that volume instead of, as in strict reality, being concentrated in a small fraction of it.
>
> Batchelor (1967: 4–5)

In some cases (as in this example of an explicit idealization), a stipulation remains just that, despite the success of the theory that takes this as a starting point.

Quine's assumption is that such idealizations will not remain in our best scientific theories. In his view, explicit idealizations such as these can be rephrased as claims about the behavior of actual systems as complicating features are minimized:

> When one asserts that mass points behave thus and so, he can be understood as saying roughly thus: that particles of given mass behave the more nearly thus and so the smaller volumes. When one speaks of an isolated system of particles as behaving thus and so, he can be understood as saying that a system of particles behaves the more nearly thus and so the smaller the proportion of energy transferred from or to the outside world.
>
> Quine (1960: 249)

However, as Maddy notes, this strategy only works for certain kinds of idealizations, which she calls "idealizations by causal isolation," where "what happens in ideal circumstances can be extrapolated from what happens in real circumstances by gradually minimizing the disturbing causal factors" (Maddy 1997: 144). The "continuum hypothesis" of fluid dynamics is not of this character:

> On Quine's principles, a claim about a continuous ideal fluid would be replaced by, or translated as, a claim about what happens to actual fluids as they approximate ever more closely to ideal fluids, that is, I suppose, as their molecules become ever more tightly packed together, approximating continuous matter. But this is all wrong. If the molecules of an actual fluid are packed tightly enough, it stops being a fluid, and even if it didn't, the best we could approximate in this way would be density, not full continuity. The real point is that fluid dynamics isn't more applicable to one fluid than another, depending on how closely that fluid approximates a continuum; rather, it provides a workable account of any fluid.
>
> Maddy op. cit.: 145

What happens to the Quinean picture of confirmation if we assume, as Yablo puts it, that some theoretical falsehoods (idealizations, metaphors, and the like), "like the poor. . . will be with us always" (Yablo 1998: 245)?

If even our best theories contain assumptions (such as the continuum hypothesis) whose literal truth we ought not accept, then Quine's holistic picture of confirmation cannot remain. Despite Quine's protestations, the example of apparently indispensable idealizations suggests that there does seem to be a distinction to be made between *mere* stipulations, and well confirmed theoretical hypotheses. And this raises the question, "How can we know whether a stipulation is to be taken as a *mere* stipulation, or to be considered confirmed by theoretical successes?" What is it that pushes a hypothesis from convenient stipulation to confirmed truth?

In my book (Leng 2010), I argue for an "explanatory" approach to the mere convention/literal truth distinction. There I suggest that, to understand which parts of our scientific theories should be considered confirmed by our theoretical successes, we should look at our best explanations *of those successes*. If the best explanation of the success of a theory requires us to assume the real existence of the objects characterized by that theory, then we ought to believe in those objects. And the best explanation will not always require such an assumption: there will be times when we can explain a theory's success on the assumption that its posits are mere fictions (the "continuous fluids" example is one such: here, we explain the theory's success by pointing out, as Batchelor does, that real fluids "move *as if* they were continuous," so that a comparison with the fictional fluids of our ideal models is apt).

It is here that the causal efficacy that Benacerraf (Benacerraf 1973) focuses on comes into play. For if a theoretical posit is assumed to play a causal role, it is very difficult to explain the success of the theory in which it is posited on the assumption that the object posited is a mere fiction. *Fictional* $\alpha$-particles do not leave tracks in cloud chambers. So when, in the context of a predictively successful theory, an object is posited and assumed to play a particular causal role, this gives us good reason for thinking that the best explanation of that theory's success involves the real existence of the object posited.

Clearly, mathematical posits do not play a causal role in our theories, so our explanation of their contribution to successful scientific theories cannot appeal to such a role. Instead, accounts of the role played by mathematical posits often focus on their *representational* role, in allowing us, in Joseph Melia's words, "to make more things sayable about concrete objects" (Melia 1998: 70–71). But this is a role

that I (and others) have argued could be played by merely fictional theoretical posits (Leng 2010), so if our explanation of the success of mathematically-stated scientific theories requires us only to attribute a successful representational role to the mathematical posits in those theories, then this will give us no reason to consider the existence of the objects posited to be confirmed by that success.

This suggests a way of thinking about the epistemic challenge raised by the acausal nature of mathematical objects that remains even in a Quinean setting, where there is no immediate causal link requirement on our knowledge of objects. The challenge is to find some theoretical role played by mathematical posits in our successful theories that could not be played equally well by *mere* stipulations (mere fictions). They do not play a causal role, so that particular route to realism is blocked. If the path from stipulation to truth is via the confirmation afforded to stipulations when they become embedded in our successful theories, we need to find some feature of our mathematical stipulations that earns them credit for that success (and that would not account for that success were the stipulated objects *mere* fictions). Jody Azzouni raises something like this challenge in his book (Azzouni 1994), where he too considers the question of why Benacerraf's knowledge problem isn't simply solved by Quinean epistemology. Azzouni focuses there on the role played by mathematical posits in the context of the practice of working mathematicians, rather than in empirical theorizing, and points out that mathematical objects do not seem to feature in the reasons mathematicians have for believing their theorems. However, Azzouni's title for the challenge he raises, the "epistemic role puzzle," remains apt when we turn to the practice of empirical science. In the context of considering mathematical practice, Azzouni asks, "What exactly is it in traditional mathematical practice that mathematical objects do?", and we may similarly ask, "What exactly is it, in empirical science, that mathematical objects do?" (that couldn't, we may add, be done by merely fictional posits).

*Do* mathematical posits make any other kind of contribution to our theoretical successes aside from enabling us to describe fundamentally physical processes? And if so, does the successful use of mathematical posits in such contexts require us to consider the objects posited as real? In recent years, Quinean platonists such as Baker (2005, 2009) and Colyvan (2002, 2010) have argued that mathematical posits also sometimes play an *explanatory* role in empirical science, and this is a role that cannot be played by mere fictions. In my own work, I have responded to this challenge on behalf of nominalists by arguing that, though Baker and Colyvan are right that mathematics does sometimes do genuine explanatory work, the way in which mathematics explains (through "structural explanations" where physical systems are seen as (approximate) instantiations of mathematically described structures) does not require the existence of any *mathematical* objects, only physical systems mathematically described (Leng 2012). But however this debate pans out, I take it that this kind of consideration of the precise role played by mathematics in empirical science is the appropriate place to focus if we wish to learn the lessons of both Benacerraf and Quine.

## References

Azzouni, J. (1994). *Metaphysical myths, mathematical practice: The ontology and epistemology of the exact sciences*. Cambridge: Cambridge UP.

Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind, 114*(454), 223–238.

Baker, A. (2009). Mathematical explanation in science. *The British Journal for the Philosophy of Science, 60*(3), 611–633.

Batchelor, G. K. (1967). *An introduction to fluid dynamics*. Cambridge: Cambridge UP.

Benacerraf, P. (1973). Mathematical Truth. *The Journal of Philosophy*, *70*(19), 661–679.

Colyvan, M. (2002). Mathematics and aesthetic considerations in science. *Mind, 111*(144), 69–74.

Colyvan, M. (2010). There is no easy road to nominalism. *Mind, 119*(474), 285–306.

Einstein, A. (1920). *Relativity: The special and the general theory*, (R. W. Lawson, Trans.). London: Methuen & Co Ldt.

Field, H. H. (1989). *Realism, mathematics, and modality*. Oxford: Basil Blackwell.

Gödel, K. [1947] (1990). What is Cantor's continuum problem? In S. Feferman (Ed.-in-chief), *Collected works*: *Publications 1938–1974* (Vol. II, pp. 515–525, 176–187). Oxford: Oxford UP.

Gödel, K. [1964] (1990). What is Cantor's continuum problem? (Revised and expanded version of Gödel [1947] 1990). In S. Feferman (Ed.-in-chief), *Collected works*: *Publications 1938–1974* (Vol. II, pp. 259–273, 254–270. Oxford: Oxford UP.

Hilbert, D. [1926] (1983). *On the infinite*. P. Benacerraf & H. Putnam, Eds. (E. Putnam & G. J. Massey, Trans., [1964] (1983), pp. 183–201.

Leng, M. (2010). *Mathematics and reality*. Oxford: Oxford UP.

Leng, M. (2012). Taking it easy: A response to Colyvan. *Mind, 121*(484), 983–995.

Maddy, P. (1992). Indispensability and practice. *The Journal of Philosophy, 89*(6), 275–289.

Maddy, P. (1997). *Naturalism in mathematics*. Oxford: Oxford UP.

Melia, J. (1998). Field's programme: Some interference. *Analysis, 58*(2), 63–71.

Quine, W. V. O. [1951] (1953). Two dogmas of empiricism. In *From a logical point of view: 9 logico-philosophical essays* (pp. 20–46). Cambridge, Massachusetts: Harvard UP.

Quine, W. V. O. (1960). *Word and object*. Cambridge: The MIT Press.

Quine, W. V. O. [1935] (1976). Truth by convention. In *The ways of paradox and other essays* (Revised and enlarged edition, pp. 77–106). Cambridge, Massachusetts: Harvard UP.

Quine, W. V. O. [1954] (1977). Carnap and logical truth. In *The ways of paradox and other essays* (2nd ed., pp. 107–132). Cambridge, Massachusetts: Harvard UP.

Russell, B. S. (1919). *Introduction to mathematical philosophy*. London: George Allen and Unwin.

Yablo, S. (1998). Does ontology rest on a mistake? *Aristotelian Society, Supplementary, 72*, 229–261.

# Chapter 8
# Identification and Transportability: Another Moral for Benacerraf's Parable of Ernie and Johnny

**Philippe de Rouilhan**

In his classic 1965 article "What Numbers Could Not Be" Benacerraf recounts the parable of Ernie and Johnny, those two children to whom their respective father-tutors presented the terms of Neumann's progression and those of Zermelo's progression as being *the* natural numbers, and who come into conflict when it comes to knowing whether or not 3 belongs to 17. From this, Benacerraf derives the ontological moral that only the *structure* of progression of the so-called natural numbers counts, while the natural numbers *themselves* do not exist. He says almost nothing about an ontologically neutral, methodological moral that would enable Ernie and Johnny to "*identify*" their respective progressions with one another, all the while collaborating harmoniously on the development of arithmetic as a theory of progressions. The aim of this paper is to present such a moral. I shall do so in terms essentially originating in the crossing of Tarski's *first*[1] theory of models, outlined in his 1936 article on logical consequence, with Bourbaki's theory of

I am grateful to Fabrice Pataut for inviting me to contribute to this collection in honor of Paul Benacerraf, the kind of offer that is not to be declined. Some of the ideas contained in this paper were intially presented, on the occasion of the *Journée* at which the very generous organizers, Pierre-Yves Quiviger, Pierre Wagner and Anna Zielinska, had invited me to give the closing talk, on December 14, 2013 at the Institut d'Histoire et de Philosophie des Sciences et des Techniques (IHPST), http://www.ihpst.cnrs.fr/activites/colloques/autour-du-travail-de-philippe-de-rouilhan-la-question-de-la-verite; and then at the seminar "De Galilée à Weil en passant par Abel and Hilbert," on June 2, 2015 at the Institut Henri Poincaré (IHP). In more recent exchanges, Serge Bozon and François Rivenc convinced me to clarify certain points. I hope I have met their expectations. At last, thank God, Claire O. Hill transmuted all the stuff, but for a few last-minute corrections, into (real) English.

Ph. de Rouilhan (✉)
IHPST (Paris 1-Panthéon-Sorbonne/ENS/CNRS), Paris, France
e-mail: rouilhan@orange.fr



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_8

structures as set forth in 1957 in the *first* edition of chapter 4 of Book 1 of the *Éléments*. I shall consider four methods, one after the other.

The first method will come to mind naturally: Ernie and Johnny have but to limit themselves to statements whose truth-value is invariant through the class of the progressions, but certain results, easy to obtain, will reveal the worrying logical complexity of the invariance in question. Thus one will be led to advise Ernie and Johnny to fall back on a second, fortunately simple, method: the *formal* axiomatic method. Taking a closer look, however, it will turn out that Ernie and Johnny should meet a certain requirement of *structurality*, explained here in terms of a revised version of Bourbaki's notion of *transportability*, and therefore move on to a third method, which one could call the *structural* axiomatic method. The concept of transportability, however, will prove to carry with it the same signs of logical complexity as the invariance of truth-value through the progressions. Ernie and Johnny could, therefore, just as well resort to a fourth and final method by going back to the Neumann and Zermelo progressions so as to uphold the requirement of transportability from that point on.

By highlighting the logical complexity of the concept of transportability, perhaps I will have added to the reasons adduced by those who doubt that Bourbaki's theory of structures is adequate for the rational reconstruction of a properly *structural* mathematics, or even that set theory is an adequate framework for such a reconstruction. I shall have, however, held back from leading our two innocent boys down the heathen path.

## 8.1 Benacerraf's Parable and Its Ontological Moral

Thus, Benacerraf presented two young children, Ernie and Johnny, who were taught arithmetic not as a particular subject but as a mere chapter of set theory.[2] For Ernie, natural numbers were defined *à la* Neumann: $0 =_{df} \varnothing$, $1 =_{df} 0 \cup \{0\}$, $2 =_{df} 1 \cup \{1\}$, …, $n + 1 =_{df} n \cup \{n\}$, …; and for Johnny, they were defined *à la* Zermelo: $0 =_{df} \varnothing$, $1 =_{df} \{0\}$, $2 =_{df} \{1\}$, …, $n + 1 =_{df} \{n\}$, …. And each one took the definition literally, as expressing what natural numbers are, namely this, or that, and no more or less than that. Everything was fine until one day, comparing their ideas, they realized with amazement that they did not agree about everything, for instance about the question as to whether or not $3 \in 17$. For Ernie the answer was yes, for Johnny no. They soon detected the origin of their disagreement, namely they had not been taught the definition of natural numbers in the same way. At least one of the two definitions, they thought, had to be wrong.

Benacerraf settled the matter in his own distinctive way, namely by pointing out that neither definition, nor any other of the same kind, provides any ground justifying its claim to prevail over its rivals when it comes to stating what natural numbers "themselves" are. Indeed, natural numbers are not really definite sets at all —neither, more generally, are they objects at all—, they are nothing at all beyond the place they hold, the part they play, the function they fulfill, the purpose they

serve within the *progression* where they occur. Those poor children, victims of the "militant logicists" they had had as their private tutors (in fact, their respective fathers!) were arguing about an illusory problem. With the eliminative version of a structuralist philosophy of arithmetic, they will (kill the father and) find their peace again.

I shall draw another lesson from that parable, one that is quite simply methodological and indeed ontologically neutral. Whatever the ontological status of natural numbers could be, whatever they could or could not be, it would suffice to explain to our two innocent boys that, in the guise of a "definition," each one was only given an *example* of progression. These examples were borrowed from Ernst Zermelo and John von Neumann, whose only aim, when it came to the "definition" of natural numbers, certainly amounted to no more than giving examples of progression.

Let me be more specific. Ernie was given Neumann's progression, $\langle \mathrm{N};\ 0,\ \mathrm{s}\rangle$,[3] with

$\mathrm{N} =_{\mathrm{df}}$ the smallest set, $y$, such that $\varnothing \in y$ and, for every $x \in y$, $x \cup \{x\} \in y$;
$0 =_{\mathrm{df}} \varnothing$;
$\mathrm{s} =_{\mathrm{df}}$ (the graph of) the function $x \mid\!\rightarrow x \cup \{x\}$ of N into N.

By the same token, Johnny was given, *with the same notation*, Zermelo's progression, $\langle \mathrm{N};\ 0,\ \mathrm{s}\rangle$, with

$\mathrm{N} =_{\mathrm{df}}$ the smallest set, $y$, such that $\varnothing \in y$ and, for every $x \in y$, $\{x\} \in y$;
$0 =_{\mathrm{df}} \varnothing$;
$\mathrm{s} =_{\mathrm{df}}$ (the graph of) the function $x \mid\!\rightarrow \{x\}$ of N into N.

One will appreciate the chiasma of first names[4] and also, more seriously, the ambiguity of notation. Ernie and Johnny argued about whether $3 \in 17$. Asked in this way, the question is ambiguous, and the first thing to do is to have them clear up the ambiguity by noting $\langle \mathrm{N_N};\ 0_{\mathrm{N}},\ \mathrm{s_N}\rangle$ Neumann's progression, and $\langle \mathrm{N_Z};\ 0_{\mathrm{Z}},\ \mathrm{s_Z}\rangle$ Zermelo's progression. The situation may then be explained straightforwardly. For Ernie the question was actually whether $3_{\mathrm{N}} \in 17_{\mathrm{N}}$ (and the answer was yes); for Johnny, whether $3_{\mathrm{Z}} \in 17_{\mathrm{Z}}$ (and the answer was no).

There is a second, incomparably more delicate, thing to teach our two ingenuous boys: we should tell them that, and why, should they want to do arithmetic together, in the modern sense of progression theory and in an harmonious manner, without either of them having to give up the poisoned gift he had received from his father-private tutor (incompatible definitions and ambiguous notation), *they can do so indeed* on the necessary and sufficient condition that they abide by a *certain requirement*. No doubt, Benacerraf is well aware of the existence of such a requirement, and what he says is important. When dealing with explications of the notion of natural number that can be provided by the means of any given particular progression, like, for instance, Neumann's or Zermelo's, he states:

> There is no way connected with the reference of number words that will allow us to choose among them, *for the accounts differ at places where there is no connection whatever between features of the accounts and our uses of the words in question*.
>
> Benacerraf (1965: 62)

In a nutshell, in order to work together harmoniously in arithmetic, Ernie and Johnny only have to limit themselves to matters that are relevant to arithmetic, either as usually practiced for its own sake, or applied. The problem is that Benacerraf does not provide any criterion that our budding arithmeticians could turn to, while this is just what they need. Let me try to do it on his place.

## 8.2 First Attempt: Ernie and Johnny Can Keep to the Genetic Method Provided They Limit Themselves to Statements Whose Truth-Value Is Invariant Through the Progressions

8.2.1 Have we not all been Ernie and Johnny, if not for the natural numbers, either presumed to be familiar or reintroduced at fresh cost as verifying the Peano axioms, at least for the successive extensions of the notion of number beyond the natural numbers: integers, rationals, reals, complex numbers, and other quaternions? At each stage, our own mentors, explicitly defining the new numbers in terms of the foregoing ones, gave us definitions of the same kind as those of which, unbeknownst to them, our two children-guinea pigs were deprived of the instructions sheet. For example, the reals were explicitly defined in terms of the rationals as one or another 19th century mathematician had defined them (Weierstrass, Cantor, Dedekind, or someone else), with an order relation and two algebraic operations that made their set into a complete ordered field. But, even imagining one started with the same rationals, these definitions did not lead to the same reals, any more than Neumann's and Zermelo's definitions lead to the same natural numbers.

And yet, no, there was no Ernie or Johnny among us, because, as when learning a language by the direct method, we later learned, as we went along, to separate the grain from the chaff, i.e., the questions, assertions and denials that must be retained from those that must be thrown out. We no more had an instruction sheet than Ernie and Johnny did, but never the slightest mishap comparable to theirs either. Left to their own devices, stuck with their trick definition, our two lost children would have needed some explanation in due form so as not to head for disaster. It is this explanation that I would like to give them. Benacerraf's parable is a far-fetched tale precisely where it makes one think the most.

8.2.2 Hilbert called the method of introducing the concept of number by successive extensions starting from the natural numbers the *genetic* method (Hilbert 1900). We can recognize in Zermelo's or Neumann's explicit definition of natural numbers a regressive extension, so to speak, of this method to the natural numbers themselves, now constructed in purely set-theoretical terms.

The problem is not only one of finding a method allowing Ernie and Johnny to develop arithmetic while avoiding any conflict and all the while retaining the original definitions and notations. A third child could join the party, who would work under conditions comparable to those of Ernie and Johnny, and for whom the natural numbers would have been defined: $0 =_{df} \varnothing$, $1 =_{df} \mathfrak{P}(0)$, $2 =_{df} \mathfrak{P}(1)$, …, $n + 1 =_{df} \mathfrak{P}(n)$, …, or for whom, more precisely, the progression of natural numbers, $\langle N; 0, s\rangle$, would have been defined by the following:

$N =_{df}$ the smallest set, $y$, such that $\varnothing \in y$ and, for every $x \in y$, $\mathfrak{P}(x) \in y$;
$0 =_{df} \varnothing$;
$s =_{df}$ (the graph of) the function $x \mapsto \mathfrak{P}(x)$ of N in N.

Ernie and Johnny may and indeed must look for the only properties of their progression that are common to all the progressions, whether or not definable. But it would be doing things backwards to begin by proving that *their progression possesses such and such a property*, so as to verify afterwards that *all the progressions possess it*, and, were this not the case, to eliminate it. The right method would be to begin by verifying that the property is indeed *arithmetically relevant* in the sense that it is either *possessed by all the progressions* ***or is not by any***, and then, in such a disjunctive case, to try to show that it is *possessed indeed by their progression*.

8.2.3 To explain more precisely under what necessary and sufficient condition Ernie and Johnny may stick to the original, ambiguous notations and incompatible definitions, the explanation itself must obviously be devoid of ambiguity. The distinctive subscripts 'N' and 'Z' introduced above (§1) are there for that purpose.

A *progression* in general can be defined, for example, in reference to Zermelo's progression, namely as being an *object of the form* $\langle U; V_1, V_2\rangle$ *such that* (i) U *is a set,* $V_1 \in U^5$ *and* $V_2 \in \mathfrak{F}(U, U)$, *and* (ii) *there is a bijection f*: $U \to N_Z$ *such that, on the one hand,* $f(V_1) = 0_Z$ *and, on the other hand, calling g the canonical extension of f to sets* $\mathfrak{F}(U, U)$ *and* $\mathfrak{F}(N_Z, N_Z)$, $g(V_2) = s_Z$.[6] Condition (i) says that the object $\langle U; V_1, V_2\rangle$ is a *structure*[7] of a certain kind, and condition (ii) that this structure is *isomorphic* to Zermelo's progression. The class of progressions thus defined is not dependent on the progression chosen as reference in the definition.

We shall call

- NA ("Neumann arithmetic") [ZA ("Zermelo arithmetic"), resp.] the extension[8] of ZFC obtained by adding the constants '$N_N$,' '$0_N$,' '$s_N$' ('$N_Z$,' '$0_Z$,' '$s_Z$,' resp.) and the definitional axioms correlative to their definition (see §1, penultimate paragraph);
- $\mathscr{L}$(NA) [$\mathscr{L}$(ZA), resp.] the language of NA (ZA, resp.) used by Ernie (Johnny, resp.) to talk about Neumann's progression (Zermelo's progression, resp.).[9]
- GA ("generic arithmetic," so to speak) the extension of ZFC obtained by adding constants 'N,' '0,' 's,' now acting as *parameters*,[10] and the axiom according to which $\langle N; 0, s\rangle$ is a progression;
- $\mathscr{L}$(GA) the language of GA, at our disposal to speak about the *generic* progression $\langle N; 0, s\rangle$.

A statement of $\mathscr{L}$(GA) is said to be *true in* (*false in*, resp.) a progression, $\langle U; V_1, V_2\rangle$,—or, more generally, a structure, $\langle U; V_1, V_2\rangle$, such that $V_1 \in U$ and $V_2 \in \mathfrak{F}(U, U)$—if, and only if, it becomes plainly *true* (*false*, resp.) when the parameters 'N,' '0,' 's' take U, $V_1$, $V_2$, respectively, as values. As for the notion of *truth* (*falsehood*, resp.) *simpliciter* appealed to here, as opposed to that of *truth in* (*falsehood in,* resp.) and not to be confused with demonstrability (refutability, resp.), it is a clear enough notion since Tarski (see Tarski 1933) for us to be allowed to use it without further explanation, or just about. To rein in the trouble and confusion sometimes accompanying the use of a word as banal as "true," may it suffice here to draw attention to this, which strikes at the heart of the concept of truth. In a quite general way, if an English statement, **A**, contains neither the truth predicate nor any related predicate, *quoting* this statement to assert its truth with the help of a second statement (⌜**A** is true⌝), or asserting what **A** expresses by *using* **A** *itself* without speaking of its truth (see endnote 12) amount to the same thing. To a certain extent, one can thus forego the truth predicate, but, if the truth predicate is applied to a statement that is merely *described*, as in "The first statement of the Vulgate is true," or is left *indeterminate*, as in "The logical consequences of a true statement are true," one cannot do this.

On the basis of all this, we can reformulate in a precise way the method that Ernie and Johnny should respectively follow. They must only take an interest in statements of $\mathscr{L}$(NA) [$\mathscr{L}$(ZA), resp.] *of which the counterpart statement of* $\mathscr{L}$(GA) [obtained by substituting 'N,' '0,' 's' for '$N_N$,' '$0_N$,' '$s_N$' ('$N_Z$,' '$0_Z$,' '$s_Z$,' resp.), respectively] is *true in every progression or is in none*, in other words, *whose truth-value is invariant through the class of progressions*, or, more briefly, is *tv-invariant through the progressions*.

Ernie and Johnny were right to affirm that $3 \in 17$ and that $3 \notin 17$, respectively, *in the sense in which they understood it*, but it is just that, from that point on, they must consider these truths, which are not tv-invariant through the progressions, to be truths *irrelevant to arithmetic*, truths *that do not count*. If they have understood correctly that the arithmetic in question is no longer the theory of *natural numbers*, but the theory of *progressions*, or of *generic progression*, they are, in principle, sheltered from any irrelevance comparable to that which led them to contradict one another.

8.2.4 To apply the method in question, the ideal would be for the tv-invariance of a statement of $\mathscr{L}$(GA) through the progressions to be easily recognizable, as, for example, Fermat's last theorem or Goldbach's conjecture, even though discovering its truth-value, presuming that to be possible, could present the greatest difficulties, as in the past for the theorem, and still today for the conjecture. Unfortunately, it may happen that this is not the case.

For the sake of accuracy, let us call $GA_+$ the extension of GA obtained by adding the *elementary syntax of* GA,[11] and $GA^+$ the extension of $GA_+$ obtained by adding what, by abuse of language, may be called a *recursive definition* of *truth in a progression* for the statements of $\mathscr{L}$(GA).[12] This theory is strong enough to

meet all our needs, and weak enough not to overflow the set-theoretical framework in which GA was itself constructed.[13] For our present purposes, it has everything in its favor.

For any statement, **A**, of $\mathscr{L}$(GA), let me call ♯**A** the universal closure of the result of substituting the variables 'U,' '$V_1$,' '$V_2$,' for 'N,' '0,' 's,' respectively, in **A**, which is the canonical expression in $\mathscr{L}$(ZFC) of the truth of **A** in all the progressions; and let me call ♭**A** the statement ⌜♯**A** or ♯(non **A**)⌝, which is the canonical expression in $\mathscr{L}$(ZFC) of the tv-invariance of **A** through the progressions.[14]

**PROPOSITIONS** — (i) *There is a statement,* **C***, of* $\mathscr{L}$(GA) *such that* ♭**C** *is as difficult to decide in* ZFC *as Goldbach's conjecture;* (ii) *there is a statement*, **C***, of* $\mathscr{L}$(GA) such that ♭**C** *is undecidable in* ZFC (*barring inconsistency*).

PROOFS.—(i). Let there be

- **A** a statement of $\mathscr{L}$(GA) *true in every progression;*
- **B** a statement of $\mathscr{L}$(GA) *non-tv-invariant through the progressions;*
- **Goldbach** the statement in $\mathscr{L}$(GA) of Goldbach's conjecture;
- **C** the statement ⌜(**Goldbach** and **A**) or (not **Goldbach** and **B**)⌝ (**C** is our statement!)

If **Goldbach** is true in all the progressions, then so is **C** and **C** is therefore tv-invariant through the progressions. And if **Goldbach** is false in all the progressions, then **C** has the same truth-value as ⌜not **Goldbach** and **B**⌝ in all the progressions, which has the same truth-value as **B** in all the progressions, which is not tv-invariant through the progressions, and therefore **C** is not tv-invariant through the progressions. We have just easily proved ⌜**C** is tv-invariant through the progressions ⇔ **Goldbach** is true in all the progressions⌝, i.e., ⌜♭**C** ⇔ ♯**Goldbach**⌝ in $GA^+$, without invoking the properly arithmetical axiom of GA. By choosing **A** = 'card(N) = card($N_Z$),' **B** = 'N = $N_Z$' and by eliminating the truth predicate of the proof, one easily obtains a proof of ⌜**C** ⇔ **Goldbach**⌝ in ZFC. It is finally as difficult to decide **C** in ZFC as it is to decide **Goldbach**.

(ii). Let us assume the consistency of ZFC (and, therefore, of GA, which is consistent relatively to it), and let us call **Gödel** a statement of $\mathscr{L}$(GA) that is undecidable in GA. By replacing '**Goldbach**' by '**Gödel**' in the beginning of the proof of (i), we deduce ⌜♭**C** ⇔ ♯**Gödel**⌝ from the hypothesis of consistency in $GA^+$ without invoking the properly arithmetical axiom of GA. By then choosing **A** and **B** as we did in the proof of (i), and by eliminating the truth predicate, we derive ⌜**C** ⇔ **Gödel**⌝ from the hypothesis of consistency in ZFC and, finally, the undecidability of **C**.

8.2.5 The proposition (ii) and its proof would in fact still hold if we replaced ZFC with any extension whatever obtained by adding an effective[15] set of purely set-theoretical new axioms. A corollary to this remark is that the set of statements of $\mathscr{L}$(ZFC) canonically expressing the tv-invariance through the progressions of a statement of $\mathscr{L}$(GA) tv-invariant through the progressions is neither effective, nor even effectively enumerable. *The same obviously goes for the set of statements*

*invariant through the progressions themselves.* We wish we had been able to say to Ernie and Johnny something like: "You can safely drop your distinctive subscript ('N' or 'Z') if, and only if, you limit yourselves to statements of your language [$\mathscr{L}$(NA) or $\mathscr{L}$(ZA)] whose counterparts in $\mathscr{L}$(GA) are the statements of a certain *sub-language* (in a sense to be specified) of $\mathscr{L}$(GA), or at least constitute a certain *effective subset* of the set of statements of $\mathscr{L}$(GA)." But, we have just seen that this is far from being the case. Let us repeat this: *the set of statements of* $\mathscr{L}$(GA) *tv-invariant through the progressions is neither effective, nor even effectively enumerable*.

For want of better, we can advise Ernie and Johnny to go to pains to establish a battery of criteria of tv-invariance through the progressions that will allow them, as arithmetic develops out of their progression, to make sure without too much difficulty that the statements they are interested in have as their counterparts in $\mathscr{L}$(GA) statements that are tv-invariant through the progressions. They will realize, for example—and this will be good news—, that *the set of statements of* $\mathscr{L}$(GA) *tv-invariant through the progressions is stable for the Boolean operations*.[16] If statements of $\mathscr{L}$(GA), **A**, **B**, **C**, …, are tv-invariant through the progressions, then so are ⌜not **A**⌝, ⌜**A** or **B**⌝, ⌜**A** and **B**⌝, ⌜**A** ⇒ **B**⌝, ⌜**A** ⇔ **B**⌝, etc. However, they will also realize, for example—and this will be bad news—, that the *set in question is not stable for tautological deduction, based just on the use of Boolean operators*. Indeed, let **A** and **B** be two statements of $\mathscr{L}$(GA) that are, respectively, true in every progression and not tv-invariant through the progressions, and let **C** = ⌜not **A** and **B**⌝. Just like **A**, **C** is tv-invariant through the progressions (because of ⌜not **A**⌝), while **B** is not although it is deducible from **C**.

To proceed further, Ernie and Johnny will first of all have to consider a property more general than that of tv-invariance through the progressions of a statement of $\mathscr{L}$(GA), viz., the property, for a *formula* of $\mathscr{L}$(GA) and an assignment of values to its free variables,[17] of *being tv-invariant through the progressions for this assignment*; then to take into account successively the formulas governed by a primitive predicate ('=,' 'Ens,' '∈') or a quantifier ('∃,' '∀'), without forgetting the description operator (say the upside down iota of Peano) in case it occurs in $\mathscr{L}$(ZFC). Our little soldiers have their work cut out for them. But, once they have come up with the right criteria, they will be saved, won't they?

## 8.3 *Second Attempt: Ernie and Johnny Would Do Better to Start Using the so-Called* Formal Axiomatic *Method*

Admittedly, Ernie and Johnny are now saved, but at what cost!

Hilbert contrasted the genetic method with the axiomatic method (Hilbert 1900), which he had implemented in his *Grundlagen der Geometrie* (Hilbert 1899) and which he preferred. Was this already the *formal* axiomatic method that he would much later contrast with the *material*, *contentual* (inhaltlich)[18] axiomatic method in

the first pages of the *Grundlagen der Mathematik I* (Hilbert 1934), as he would have then people believe and as the *doxa* adamantly maintains? Be that as it may, in order to see the formal axiomatic method unhesitatingly adopted on a major scale, the historical priority must be given to van der Waerden in his famous *Moderne Algebra* (van der Waerden 1931).[19]

Instead of allowing our two heroes to go to great pains to observe the safety measures prompted, strictly speaking, by the use of the genetic method, we should rather initiate them to the formal axiomatic method, applied to arithmetic considered once again as theory of progressions, or of generic progression, and reveal to them the only advantage for the new method that can be had from the rotten definitions that their respective fathers made them swallow.

So Ernie and Johnny only have to work directly in GA, or better, in the arithmetic obtained from it by replacing its single but needlessly complicated axiom with the celebrated axioms chosen to perfection by the author of *Arithmetices*[20] *Principia* (Peano 1889). In other words, they have but to work directly in Peano arithmetic,[21] referred to here as PA,[22] whose axioms express: (A1) that 0 belongs to N, (A2) that s is (the graph of) an application of N into N, (A3) that this application is injective, (A4) that 0 does not belong to its image, and (A5) that any subset of N having 0 among its members and being stable under s is all of N.

Of course, our two young modern axiomaticians will not therefore be shielded from arithmetic *irrelevance* (see above, §1), but, *assuming that* PA *is consistent*, at least they will not risk *contradicting* one another. As for the sole genuine interest of the purely set-theoretical construction of progressions such as those of Neumann and Zermelo, it shows in ZFC the existence of progressions, and, as a corollary, that PA *is consistent relative to* ZFC. Indeed, *if* there existed a statement, **A**, of $\mathscr{L}$(PA) such that **A** and its negation were provable in PA, in other words, such that **A** and its negation were deducible in a purely set-theoretical fashion from the Peano axioms, *then*, by substituting the *definiensa* of '$N_Z$,' '$0_Z$,' '$s_Z$,' for example, for 'N,' '0,' 's,' respectively, in corresponding deductions, one would obtain deductions in ZFC of a statement and its negation from instances of Peano axioms provable in ZFC, and ZFC would therefore be contradictory.[23]

## 8.4 *Third Attempt: Ernie and Johnny Would Do Even Better to Start Using an* Axiomatic *Method that One Might Call* Structural

8.4.1 Is everything fine, now? No, not yet.

Let us go back to the conflict to which the young students have found themselves exposed through the fault of irresponsible mentors and to its solution in terms of tv-invariance through the progressions. In the beginning, we more or less told them this: "If (and only if) you limit yourselves to the statements of your language that attribute to your progression a property that is possessed either by all the

progressions or by none, i.e., if you limit yourselves to statements whose counterparts in $\mathscr{L}$(GA), or $\mathscr{L}$(PA), are tv-invariant through the progressions, you will be protected from any conflict, you will be able to drop your distinctive subscript safely. You will be able, as one says, to *identify* your respective progressions with one another, and also with any progression and with the generic progression." Then, we saw how costly this precaution would be and finally urged our arithmeticians in short pants to change methods, to adopt the formal axiomatic method for developing GA or, better, PA.

The analysis was not wrong, but it might have been somewhat obsessed with the maxim of conservation in solving problems. It remained superficial. It did not go to the heart of the matter. Right after having considered the first method with the two boys, we could have, and perhaps even should have, told them in heuristic terms: "In fact, apart from any danger of conflict, the definitions you were given without instructions for their use have their origin, their own methodological nature, in the idea that what counts in the study of a progression as such isn't, of course, all the properties of this progression, but it isn't, for that matter, all the properties that progressions have in common either. What counts is *only* those of these properties that are—in a sense yet to be specified—*structural*. You must therefore limit yourselves to statements of your language that express the possession of these properties by your progression, in other words, to those of these statements whose counterpart in $\mathscr{L}$(GA) is *structural*." And it is now time to tell them: "In fact, it is not all the properties of the generic progression that count. It is only its *structural* properties. You must therefore limit yourselves to statements of $\mathscr{L}$(PA) that are *structural*." Be that as it may of the *ontological* structuralism inaugurated by Benacerraf about arithmetic, it is in this requirement of structurality, and in this requirement only, that a rigorous idea of *methodological* structuralism arises.

8.4.2 To explain what that means nicely, freely echoing Bourbaki's theory of structures as applied to the theory of progressions, I must enrich the supply of model-theoretical notions presented above (§2.2) a little bit.

The axioms of PA naturally divide into two groups. The first two axioms:

(A1) $0 \in \mathrm{N}$
(A2) $\mathrm{s} \in \mathfrak{F}(\mathrm{N}, \mathrm{N})$[24]

are the *axioms of typification*, which say that 0 and s belong, respectively, to *echelons* N and $\mathfrak{F}(\mathrm{N}, \mathrm{N})$ of the *scale of sets* of *base* N.[25] The last three ones:

(A3) s is injective
(A4) $0 \notin \mathrm{Im(s)}$
(A5) for every $\mathrm{X} \subseteq \mathrm{N}$, if $0 \in \mathrm{X}$ and X is stable under s, then $\mathrm{X} = \mathrm{N}$

are the *axioms of specification*, which say which purely set-theoretical properties the objects N, 0, s are assumed to possess, or in which relations they are assumed to be.

An object of the form $\langle \mathrm{U};\ \mathrm{V}_1,\ \mathrm{V}_2 \rangle$, where $\mathrm{V}_1 \in \mathrm{U}$ (see endnote 5) and $\mathrm{V}_2 \in \mathfrak{F}(\mathrm{U}, \mathrm{U})$, is a *structure of interpretation* of the language $\mathscr{L}$(PA), i.e., a

structure in which $\mathscr{L}$(PA) can be *interpreted*, the *interpretation* being then that in which 'N,' '0,' 's' denote—or take as a value—U, $V_1$, $V_2$, respectively. **A** (**M**, resp.) being a statement (a set of statements, resp.) of $\mathscr{L}$(PA), such a structure is a *model* of **A** (**M**, resp.) if, and only if, **A** (the statements of **M**, resp.) is (are) *true in* this structure. The class of models of the axioms of typification is a *kind of structure*, referred to here as Γ; the class of structures of kind Γ that are models of the axioms of specification is a *species of structure*, referred to here as Σ, subordinated to the kind Γ. The structures of species Σ are none other than the progressions; the structures of kind Γ have not been assigned a common noun of their very own.

8.4.3 The essential question remains to be answered: What does the *structurality* requirement consist in? The structurality of a statemenof $\mathscr{L}$(PA) mustcertainly be explicable in terms of *tv-invariance by isomorphism through a certain class of structures*. The question is: Which one? Since there is only one class of isomorphism in the species Σ, the tv-invariance through Σ is none other than the tv-invariance *by isomorphism* through Σ, but Σ is not the class we are looking for. The explication of structurality by tv-invariance by isomorphism through Σ would underdetermine the essence of the structurality one wants to explain. It would at best only explain a *weak* form of structurality, structurality *relative* to Σ.

A preliminary remark is necessary regarding the notion of isomorphism. In mathematics, people have acquired the habit of associating the notion of isomorphism with a species of structure ("isomorphism of *ordered sets*," "isomorphism of *groups*," "isomorphism of *topological spaces*," etc.). Admittedly, this way of speaking is formally correct, but it is also somewhat misleading. It would lead us here to speak of "isomorphism of *progressions*," as if the notion of isomorphism at play was bringing about the relativization to *species* Σ all by itself, while it depends only on the *kind* Γ. Two structures of kind Γ, $\langle U; V_1, V_2\rangle$ and $\langle U'; V'_1, V'_2\rangle$, are isomorphic if, and only if, there is a bijection, $f\ U \rightarrow U'$, such that $f(V_1) = V_1'$ and, calling $g$ the canonical extension of $f$ to sets $\mathfrak{F}(U, U)$ and $\mathfrak{F}(U', U')$, $g(V_2) = V'_2$. Now, when PA is developed ab initio (within the framework already fixed of ZFC), once the axioms of typification are laid down and the kind Γ and the notion of isomorphism are thereby both determined, the structurality requirement holds starting from the first possible axiom of specification, no matter which one it is, even though it has not been laid down yet. Just like the notion of isomorphism, the notion of structurality to be explicated must depend only on kind Γ and not on one or another subordinate species. This thesis can be confirmed by noting that the development of PA naturally leads to an interest in the consistency of the axioms of specification, in their mutual independence, and, more generally, in the effect each one of them has on the theory, and therefore to a consideration, however brief, of other species of the same kind Γ, while an interest in the very same statements is still in order. These statements must be able to retain their structural nature independently of the species under consideration.

Hence the explication sought, based on the strongest possible form of tv-invariance by isomorphism, according to which *a statement of* $\mathscr{L}$(PA) *is*

*structural (relative to* $\Gamma$*) if, and only if, it is tv-invariant by isomorphism through the kind* $\Gamma$—in other words, still borrowing just as freely from Bourbaki, if, and only if, it is *transportable (through* $\Gamma$*)*. The *explicatum* of structurality is *transportability*.[26]

Any transportable statement is obviously tv-invariant through $\Sigma$, but the reverse does not hold, as the following counter-example shows. Let us use 'A1–5' to abbreviate the conjunction of axioms (A1), …, (A5), and let $\langle Z; 0, s\rangle$ be a structure of kind $\Gamma$ defined purely set-theoretically and isomorphic to the set of integers equipped with its usual zero and successor function. The statement 'A1–5 or $\langle N; 0, s\rangle = \langle Z; 0, s\rangle$' is true in the progressions, it is therefore tv-invariant through $\Sigma$; but it is also true in $\langle Z; 0, s\rangle$ without being true of any other structure isomorphic to the latter. It is therefore not transportable.

The axioms of typification of PA are transportable, since they are true, by definition, in any structure of kind $\Gamma$. So are, as they should be, the axioms of specification, as can be verified axiom by axiom. And it is in keeping with the spirit of the structural method, even though the author of the *Éléments* does not do it expressly for any *theory of species of structure as such*, to require of the theorems of the theory of progressions *as theory of the species of structure of progression* that they be transportable.

## 8.5 Fourth Attempt: Ernie and Johnny Could just as Well Go Back to the Genetic Method, Provided They Limit Themselves to Transportable Statements

Subject to the requirement of transportability, the structural axiomatic method no longer enjoys the obvious advantage the formal axiomatic method had over the genetic method. If Ernie and Johnny want to go back to the genetic method and, by studying their favorite progression, to find exactly the same transportable properties as those of the generic progression that the structural axiomatic method enables one to find, they will have to endorse *essentially the same* requirement, and it will be enough for them to do so. More precisely, they will have to keep to the statements of their language whose counterpart in $\mathscr{L}$(GA) is transportable. On this condition, and on this condition alone, will they be able to *identify*[27] their respective progressions with one another and also to *identify* them with the generic progression—i.e., they will be able to *delete their distinctive subscript* (and thus to restore the original ambiguity)—without running the risk of mutual contradiction any more than that of a divergence on the set of statements selected.

A balanced judgment can be made of the situation, which could already have been made in Sect. 8.2.3. On the one hand, it is easy to show that transportability (through $\Gamma$) has the same properties, and therefore poses the same problems as tv-invariance through $\Sigma$. On the other hand, the transportability of a statement is, as a matter of practice, more or less accessible to the intuition of anyone understanding

it. This intuition may perhaps be validated by a proof, and this proof will be all the easier to find if the suitable criteria of transportability have been established once and for all.

In the above-mentioned appendix, Bourbaki sets up some thirty criteria of transportability and finally leaves readers with a prudent, but optimistic remark:

> *Remark.* — Practice alone can teach the extent to which the identification of two [structures][28] presents more advantages than disadvantages. It is necessary in any case that when using it one does not risk writing non-transportable [formulas].[29] The criteria given in this Appendix show that the risk is most often minimal.
>
> Bourbaki (1957) [1st ed.]: Chap. 4, 69

Whether or not these fine words suffice to reassure those they reach, one could ask whether, in laying bare some not very nice properties of the notion of transportability, we have not, *volens nolens*, added a bit of new grist to the mill of those who, for completely different reasons which have arisen with the development of mathematics itself, have long doubted or denied that the Bourbachic theory of structures is adequate for the rational reconstruction of a properly structural mathematics. Most of these disbelievers eye the theory of categories, the moderates pleading for a new explication, "categorical" in style, in an adapted set-theoretical framework, and the radicals for a change of framework and a "categorical" reconstruction of all of mathematics.[30] When the day comes, our budding mathematicians will also have to take a stand on this matter and proceed, if they think it necessary, to engage in some more or less painful revision.

Until that time, let us let them fight their first battle in light of the notion of transportability and of the identifications that it authorizes: a time comes for everything, even in the case of little geniuses.

## Notes

1. Not to be confused with the *second* theory of models, codified by Tarski et al. in the 1950s, which has developed into one of the fundamental topics of contemporary mathematical logic.
2. To fix our ideas, one can identify the set theory in question to Zermelo-Fraenkel set theory with the axiom of choice (ZFC). This is what I shall do—except for a detail, which, nevertheless, is not to be neglected. ZFC is usually presented by excluding from its intended universe those *non-sets* that Zermelo called *Urelemente*. On the contrary, following Zermelo's example, I am anxious to countenance the possibility of such objects. The usual axioms must then be amended to accommodate this possibility, and it turns out that the language must contain, beyond the usual '$\in$' (dyadic predicate of membership), a second, properly set-theoretical, primitive constant, for instance 'Set' (monadic predicate of *sethood*) or '$\varnothing$' (singular term for the *empty set*).
3. $\langle N; 0, s\rangle$—definable as being $\langle\langle N\rangle, \langle 0, s\rangle\rangle$—rather than $\langle N, 0, s\rangle$, to mark the distinction between, on the one hand, the base set N, and, on the other hand, the operations 0 (zero-adic operation) and s (monadic operation) on this set.

4. I have always sensed that this chiasma was deliberate. When writing this paper, I expressed this to Fabrice Pataut, who then raised the question with Benacerraf. The answer (to Fabrice) did not disappoint me: "You have found me out! I am embarrassed to confirm that it was my idea of a small joke, and, in the context, an example of a harmless switch about which there WAS a fact of the matter regarding which was the right way."
5. If $V_1 \in U$, then U is a set and the clause that it is so is redundant and may be eliminated. In the analogous circumstances encountered later on, the corresponding clause will be eliminated.
6. The requirement that $g(V_2) = s_Z$ is equivalent to the more familiar one that, for all $x \in U$, $f(V_2(x)) = s_Z f(x)$).
7. For a definition of the general notions of structure and isomorphism in perfect agreement with the considerations of the present article, see Rouilhan (2007: 43–56), to be compared with Bourbaki (1957) [1st ed.]: chap. 4, §1 (without substantial change in the subsequent editions).
8. It will go without saying that, in going from a theory to one of its extensions, the validity of the first theory's axiom schemata and rules of inference are extended to the language of the second theory.
9. To simplify things, setting aside the definition of these constants, I consider any other possible "definition" to be only an abbreviation not involving any enrichment of the language in question.
10. Without claiming to explain the general notion, I shall say that the *parameters* are letters dealt with syntactically as constants, but semantically as variables, to the extent that their interpretation is, in a certain sense, "indeterminate, but fixed," as was formerly said. In this case, in their new, parametric role, 'N,' '0,' 's' are no longer ambiguous notations for designating $N_N$, $0_N$, $s_N$ as well as $N_Z$, $0_Z$, $s_Z$, respectively. They can now respectively designate any U, $V_1$, $V_2$ such that $\langle U; V_1, V_2\rangle$ is a progression.
11. It is a matter of the elementary study of signs, finite sequences of signs [among them, the terms, the formulas, the statements (*i.e.* closed formulas)], and finite sequences of finite sequences of signs (among them, the demonstrations), independently of their meaning.
12. It is in reality a matter of an (explicit) definition of *truth in a progression* for the *statements* of $\mathscr{L}$(GA) in terms of a more general notion, which concerns the *formulas* of $\mathscr{L}$(GA) and is *defined by recursion on their length*, viz., the notion of *truth of a formula in a progression for an assignment of values to its free variables*. [A statement being a formula without free variables, its truth in a progression can be explicitly defined as truth in this progression for every (or some) assignment of values to its free variables.] The "recursive definition of truth in a progression" in question can be considered adequate insofar as, for any statement *explicitly given*, **A**, of $\mathscr{L}$(GA), the statement "for any progression $\langle U; V_1, V_2\rangle$, **A** is true in $\langle U; V_1, V_2\rangle \Leftrightarrow A[U, V_1, V_2]$," where 'A[U, $V_1$, $V_2$]' is an abbreviation for the formula obtained from **A** by substituting 'U,' '$V_1$,' '$V_2$,' for 'N,' '0,' 's,' respectively, is provable in $GA^+$ without resorting to the properly arithmetical axiom of GA. This is the case, e.g., of the statement "for

every progression $\langle U; V_1, V_2\rangle$, 'for every $x \in N$, $s(x) \neq x$' is true in $\langle U; V_1, V_2\rangle$ $\Leftrightarrow$ for every $x \in U$, $V_2(x) \neq x$." [Compare with the remark "striking at the heart of the concept of truth" made above (§ 2.3).].

13. This means that, like $\mathscr{L}(GA)$, $\mathscr{L}(GA^+)$ is an extension of $\mathscr{L}(ZFC)$ obtained by adding constants each one of which is a singular term, a predicate, or a functor, without affecting the variables [it goes without saying that, like those of GA, the sets formed respectively by the terms, the formulas, and the demonstrations of $GA^+$ are *effective* (about this notion, see footnote 16)]. From this point of view, the only notable difference between $\mathscr{L}(GA)$ and $\mathscr{L}(GA^+)$ is that the additional constants of $\mathscr{L}(GA)$ are all singular terms, while one of the additional constants of $\mathscr{L}(GA^+)$ is a predicate (it is the predicate of *truth of a formula in etc*. mentioned above, footnote 12).
14. The equivalences $\ulcorner\mathbf{A}$ is true in all the progressions $\Leftrightarrow \flat\mathbf{A}\urcorner$ and $\ulcorner\mathbf{A}$ is tv-invariant through the progressions $\Leftrightarrow \#\mathbf{A}\urcorner$ are provable in $GA^+$ in a purely set-theoretical fashion, without appealing to the properly arithmetical axiom of GA (see endnote 12).
15. The notions of *effectivity* and *effective denumerability* are informal. The idea here is that a set of expressions (or its characteristic function) is *effective* if, and only if, there is an *effective procedure* enabling one, for any (explicitly given) expression, to settle in a finite number of steps the question whether or not it belongs to the set in question; and that a set of expressions is *effectively denumerable* if, and only if, there is an *effective procedure* making it possible to draw up a finite or indefinite list of elements (each explicitly given in its turn) of this set. (One can persuade oneself that the effectivity of a set of expressions implies its effective denumerability, but not vice versa.) As for the notion, also informal, of *effective procedure*, several formal explications of it were given in the 1930s, the most convincing being that of Turing, which however all proved equivalent to one another. *Church's thesis* states that they are adequate. The general theory of any one of the *explicata* can naturally be had within the framework of ZFC.
16. Operations also called *truth-functional*. It is a matter of operations of negation, disjunction, conjunction, etc.
17. It goes without saying that these variables may be assigned absolutely any values, thus without any restriction whatsoever involving N, 0, and/or s.
18. Instead of contrasting along with Hilbert the "formal" with "material" methods, people sometimes contrast the "modern" with "traditional" methods—a mere matter of words.
19. Which, being a matter of *numbers*, does not, however, keep the author—after having presumed the natural numbers to be familiar and having recalled the Peano axioms—from successively introducing the (relative) whole numbers, the rationals, the reals, the complex numbers and the quaternions by the genetic method.
20. "*Arithmetices"* instead of "*arithmeticae*," in "*Latino sine flexione*," the *Interlingua* invented by Peano.

21. Which it would be fairer to call "Dedekind-Peano arithmetic"; see Dedekind (1888).
22. Not to be confused with the homonymous theory ("first order Peano arithmetic," "PA"), which logicians have rightly taken a great deal of interest in, but the expressive and demonstrative capacities of which are extremely limited compared to the one of interest to us here.
23. In a general way, from the construction in ZFC of an ordered set, a group, a topological space, etc., one can conclude that the theory of ordered sets, group theory, the theory of topological spaces, etc. are consistant relative to ZFC. It is to be noted that the reverse does not hold. For example, the theory of sets of cardinality strictly larger than that of the denumerable and smaller than that of the continuum, which one could present in the formal axiomatic mode with 'E' as sole proper constant (playing the role of parameter), and '$\aleph_0 <$ card (E) $< 2^{\aleph_0}$' (which implies that E is a set), as sole proper axiom,—this theory is consistent relative to ZFC (Cohen 1963–1964), but one cannot for that matter construct a set in ZFC whose cardinal has this property (one can do so only if ZFC is inconsistent).
24. In general, if E and F are sets, $\mathfrak{F}(E, F)$, or $F^E$, is the set of (the graphs of) the applications of E into F.
25. The scale in question, $\mathfrak{E}(N)$, is the smallest set, X, such that $N \in X$ and, for all sets $E_1, E_2, \ldots, E_{n+1} \in X$, with $n \geq 1$, $\mathfrak{P}(E_1 \times E_2 \times \ldots \times E_n) \in X$ and $\mathfrak{F}(E_1 \times E_2 \times \ldots \times E_n, E_{n+1}) \in X$. The members of the scale $\mathfrak{E}(N)$ are its echelons. Let us say that, in this scale, the members of the echelon N are of *type* 1 and that, if the members of echelons $E_1, E_2, \ldots, E_n, E_{n+1}$ are of respective *types* $\tau_1, \tau_2, \ldots, \tau_n, \tau_{n+1}$, then the members of $\mathfrak{P}(E_1 \times E_2 \times \ldots \times E_n)$ [$\mathfrak{F}(E_1 \times E_2 \times \ldots, E_n, E_{n+1})$, resp.] are of *type* $\langle \tau_1, \tau_2, \ldots, \tau_n \rangle$ ($\langle \tau_1, \tau_2, \ldots, \tau_n \rightarrow \tau_{n+1} \rangle$, resp.). The members of N [$\mathfrak{F}(N, N)$, resp.] are therefore the objects of type 1 ($\langle 1 \rightarrow 1 \rangle$, resp.) in the scale $\mathfrak{E}(N)$. To situate an object in the scale $\mathfrak{E}(N)$, instead of saying what echelon it belongs to [for example, 0 to N and s to $\mathfrak{F}(N, N)$, see (A1) and (A2)], one could say what its "type" is (1 for 0 and $\langle 1 \rightarrow 1 \rangle$ for s). Hence the terminology of "typification," the usefulness of which, however, is only clearly apparent in the *general* theory of structures.
26. In the same way, tv-invariance by isomorphism through $\Sigma$, which explicates structurality relative to $\Sigma$, as considered at the beginning of the present Sect. 4.3, corresponds to *transportability relative* to $\Sigma$ (Bourbaki 1957 [1st ed.]: Appendix, n° 4). The *explicatum* of relative structurality is *relative transportability.*
27. See Bourbaki (1957) [1st ed.]: Chap. 4, Appendix, whose n°5 is entitled "Identifications." This appendix was purely and simply eliminated from the following editions. One would like to know exactly what kind of pressure was applied to bring about this unfortunate decision. The archives available online today do not afford an answer.
28. Bourbaki is speaking here about "sets" for reasons of context, but what holds for "sets" holds in particular for (what I call) "structures.".
29. In Bourbaki's idiosyncratic terminology: relations.

30. On the complex story of Bourbaki's relationship to the theory of categories, see Corry (1992, 1996), in particular Chaps, 7 and 8 (§ 8.5), and Krömer (2006). For a radical "categorical" solution to the Ernie and Johnny problem, see Lawvere (1964), where an elementary, "categorical" theory of sets is developed, and McLarty (1993), where this theory is applied with a view to such a resolution, whose value, whatever it may be, stems entirely from that of the theory itself.

## References

Benacerraf, P. (1965). What numbers could not be. *The Philosophical Review*, *74*, 47–73.

Bourbaki, N. (1957). *Éléments de Mathématique, Théorie des ensembles*, chap. 4 (*Actualités Scientifiques et Industrielles*, Vol. 1258), Hermann, Paris (2ème éd. 1966, without the important appendix of the 1st Ed., pp. 51–69, which was purely and simply deleted; 2nd Ed anonymous translated into English. In Bourbaki 1968, pp. 259-346).

Bourbaki, N. (1968). Elements of Mathematics, Theory of sets, in one volume, Hermann: Paris and Addison-Wesley: Boston [curiously enough, chap. 1 and 2 are translations from the 2nd, French, ed. (1960), not from the 3rd (1966)].

Cohen, P. J. (1963–1964). The independence of the continuum hypothesis. I, *Proceedings of the National Academy of Sciences, U. S. A.* (Vol. 50 (1963), pp. 1143–1148); II, *ibid.* (Vol. 51 (1964), pp. 105–110).

Corry, L. (1992). Nicolas Bourbaki and the concept of mathematical structure. *Synthese, 92*(3), 315–348.

Corry, L. (1996). *Modern algebra and the rise of mathematical structures*, series *Science Networks —Historical studies* (Vol. 17, 2nd Ed. 2004) Basel-Boston-Berlin: Birkhäuser.

Dedekind, R. (1888). *Was sind und was sollen die Zahlen*, Vieweg und Sohn, Braunschweig [translated as The nature and meaning of numbers. In W. W. Beman (Ed.) (1901) *Essays on the theory of numbers* (pp. 29–115). Chicago: Open Court; extensively revised version of Beman's translation by W. B. Ewald (Ed.) (1996). In *From Kant to Hilbert. A Source Book in the Foundations of Mathematics* (Vol. 2, pp. 787–833), Clarendon Press, Oxford].

Hilbert, D. (1899). Grundlagen der Geometrie. In *Festschrift zur Feier der Enthullung des Gauss-Weber Denkmals in Göttingen*, Teubner, Leipzig (later Ed. 1903, 1909, 1913, 1922, 1923, 1930, 1956, 1962, 1968; 1st Ed. translated into English by E. J. Townsend as *Foundations of Geometry*, Open Court, Chicago, 1902; 10th Ed. translated into English by L. Unger under the same title, La Salle: Open Court).

Hilbert, D. (1900). Über den Zahlbegriff, *Jahresberichte der Deutschen Mathematiker-Vereinigung* (Vol. 8, pp. 180–194) [translated as On the concept of number. In W. B. Ewald (Ed.) (1996) *From Kant to Hilbert. A source book in the foundations of mathematics* (Vol. 2, pp. 1089–1095), Clarendon Press, Oxford.].

Hilbert. (1934–1939). *Grundlagen der Mathematik* (Vol. 1, 1934, Vol. 2, 1939), Berlin: Springer. (2nd Ed. Vol. 1, 1968, Vol. 2, 1970); (both editions are translated into French by F. Gaillard and M. Guillaume (2001) as *Fondements des mathématiques* (Vol. 1 et 2), Paris: L'Harmattan).

Krömer, R. (2006). La 'machine de Gröthendieck' se fonde-t-elle seulement sur des vocables métamathématiques? Bourbaki et les catégories au cours des années cinquante. *Revue d'histoire des mathématiques, 8*, 119–162.

Lawvere, F. W. (1964). An elementary theory of the category of sets. *Proceedings of the National Academy of Sciences, 52*, 1506–1511.

McLarty, C. (1993). Numbers can be just what they have to. *Noûs, 27*(4), 487–498.

Peano, G. (1889). *Arithmetices principia nova methodo exposita*, Bocca, Turin [partially translated as The principles of arithmetic, presented by a new method. In J. van Heijenoort (Ed.) (1967) *From Frege to Gödel. A source book in mathematical logic 1879–1931* (pp. 85–97). Cambridge, Mass: Harvard UP; translated in its entirety as The principles of arithmetic, presented by a new method. In H. C. Kennedy (Ed.) (1973) *Selected Works of Giuseppe Peano* (pp. 101–134). London: George Allen & Unwin].

Rouilhan, Ph. de (2007). La théorie des modèles et 1' architecture des mathématiques. In P. Gochet et Ph. de Rouilhan, *Logique épistémique et philosophie des mathématiques*. Paris: Dunod, pp. 39–114, 115–118.

Tarski, A. (1933). Projecie prawdy w jezykach nauk dedukcyjnych (The concept of truth in the languages of deductive sciences), Warszawa (German Ed. as Der Wahrheitsbegriff in den formalisierten Sprachen. In *Studia Philosophica 1,* 1935, 261–405; English Ed. as The concept of truth in formalized languages. In Tarski 1956, pp. 152–278).

Tarski, A. (1936). Über den Begriff der logische Folgerung. In *Actes du Congrès International de Philosophie Scientifique, Paris 1935*, Vol. 7 (*Actualités Scientifiques et Industrielles*, Vol. 394), pp. 1–11 (translated into English as On the concept of logical consequence. In Tarski 1956, pp. 409–420).

Tarski, A. (1956). *Logic, Semantics, Metamathematics*. *Papers from 1923 to 1938*; translated by J. H. Woodger. Oxford: Clarendon Press (2nd Ed. by J. Corcoran, Hackett Publishing Company, 1983).

Waerden, B. L. van der (1931). *Moderne Algebra* (two Vol.), Berlin: Springer. (2nd Ed. 1937, 3rd Ed. 1940; translated from the German, 2nd Ed. into English by Th. J. Benac (1950) as *Modern algebra*, New York: Ungar).

# Chapter 9
# What Numbers Could Be; What Objects Could Be

**Stewart Shapiro**

Paul Benacerraf's "What Numbers Could Not Be" (Benacerraf 1965) has dominated thinking in the philosophy of mathematics for almost 50 years. There have been dozens of influential papers and books telling us what natural and real numbers could and could not be, in light of the main observations forcefully presented in that paper. Benacerraf's key observation, of course, is that there are multiple, equally good reductions of natural numbers within set theory and, for that matter, within just about any proposed foundation of mathematics.

Benacerraf (1965) was a major motivation, perhaps *the* major motivation, for various varieties of structuralism, views that see mathematics as the science of structure. According to the structuralist, what matters about natural numbers, for example, are their relations to other natural numbers. In metaphysical terms, the idea is that natural numbers do not have substantial *intrinsic* properties. They are not self-subsistent objects, on a par with, say, tables, cars, and human beings. As Benacerraf put it, "[t]o *be* the number 3 is no more and no less than to be preceded by 2, 1, and possibly 0, and to be followed by 4, 5, and so forth" (Benacerraf op. cit.: 70).

Roughly speaking, there are two varieties of structuralism. *Eliminative* structuralists hold that, due to the various features highlighted in "What Numbers Could Not Be," natural numbers are not objects (Parsons 1990). The same goes for real numbers, complex numbers, real-valued functions, and maybe even sets. We have, to modify the title of Geoffrey Hellman's (1989) book, mathematics without objects; or, to use the title of Burgess and Rosen's 1997 book, mathematics is a subject with no object. It is, I think, straightforward to interpret the final section of Benacerraf (1965) as a defense of eliminative structuralism. Michael Dummett dubbed this a "hardheaded" orientation to structuralism:

S. Shapiro (✉)
Ohio State University, Columbus, USA
e-mail: shapiro.4@osu.edu



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_9

> According to it, a mathematical theory, even if it be number theory or analysis which we ordinarily take as intended to characterize *one* particular mathematical system, can never properly be so understood: it always concerns all systems with a given structure.
>
> Dummett (1991: 296)

Benacerraf (1965) is also a major inspiration for a second, opposing view, *ante rem* structuralism, as articulated in my own *Philosophy of Mathematics: Structure and Ontology*, from 1997. With the eliminativists, I take natural numbers to be places in the natural number structure, but I take this structure to exist independently of any exemplifications it may have. Structures are like traditional universals, in that each structure is a one-over-many. In this case, however, the "many" are not individual objects, but rather systems of related objects. *Ante rem* structuralism opts for a platonic orientation to such things. As such, the natural number structure is a particular ω-sequence, the form of all ω-sequences. And, at least in some uses, numerals denote places within this structure. Natural numbers, so understood, are bona fide objects. Dummett calls this "mystical structuralism" (in Dummett op. cit.).

The two varieties of structuralism agree that there is no more to being a particular natural number than being related to other natural numbers in certain ways. One view says that, because of this, numbers are not objects; the other view insists that numbers are objects, namely places in an *ante rem* structure. One key philosophical difference between the two structuralisms concerns what it takes to be an object. This is our present topic.

Being the classic that it is, "What Numbers Could Not Be" deserves reading and re-reading, especially as new views and options are put on the philosophical map. What I propose to do here is to re-examine the third, and final, section, entitled "Way out." Our agenda is to see what views on objecthood the considerations there support, or at least suggest. I propose connections to other work in the philosophy of mathematics and the philosophy of language.

There is, I suggest, a tension between the first and third sub-sections of "Way out." The first proposes an intriguing account of identity. Extending those insights to the very notion of object gives a different "way out," a way in which natural numbers are objects after all. Moreover, if certain views in the philosophy of language are correct, numbers, so construed, are not all that different from other sorts of objects. We will consider the sub-sections in order.

## 9.1 Sub-section A: "Identity"

The new, intriguing proposal is that statements of identity are context sensitive. Benacerraf begins with a critical presentation of an aspect of Frege's position (e.g., Frege [1884] 1960), the aspect that led to the infamous Caesar problem:

> To speak from Frege's standpoint, there is a world of objects — that is, the designata or referents of names, descriptions, and so forth — in which the identity relation has free reign. It made sense for Frege to ask of *any* two names (or descriptions) whether they named the same object or different ones. Hence, the complaint at one point in his argument that, thus far, one could not tell from his definitions whether Julius Caesar was a number.
>
> Benacerraf (1965: 64)

Benacerraf then proposes an alternate view, one that denies "that all identities are meaningful." In particular, the Caesar identities are dismissed as "senseless" or "unsemantical":

> Identity statements make sense only in contexts where there exist possible individuating conditions. If an expression of the form "$x = y$" is to have a sense, it can be only in contexts where it is clear that both $x$ and $y$ are of some kind or category $C$, and that it is the conditions which individuate things *as the same C* which are operative and determine its truth-value. […] To put the point differently, questions of identity contain the presupposition that the "entities" inquired about both belong to some general category. This presupposition is normally carried by the context or theory […]. To say that they are both "entities" is to make no presupposition at all — for everything purports to be at least that.
>
> […] There are really two correlative ways of looking at the problem. One might conclude that identity is systematically ambiguous, or else one might agree with Frege, that identity is unambiguous, always meaning sameness of object, but that (contra-Frege now) the notion of an *object* varies from theory to theory, category to category […]. This last is what I am urging.
>
> Benacerraf op. cit.: 64–65, 65, 65–66

I would urge "this last" as well.

Notice, first, that this is *not* a kind of relative identity, such as that advocated by Peter Geach (Geach 1967). Benacerraf is not saying that there could be two categories $C_1$, $C_2$, such that $s$ is the same $C_1$ as $t$, but $s$ is a different $C_2$ from $t$—for the very same $s$ and the very same $t$, whatever *that* might mean. Benacerraf's thesis is that each singular term, and each variable, is associated with a particular category. Statements of identity make sense only if the items that flank the identity sign are associated with the same category. Or, in material mode, identities only make sense between entities from the same category.

As Benacerraf puts it, statements of identity have a *presupposition* that the entities in question belong to the same category. It might be noted that identity statements, on this view, pass at least one main test for presupposition: both statements of identity and their negations presuppose sameness of category. So, on this view, when the entities come from different categories, the corresponding identity statement suffers from presupposition failure, and so has no truth-value. The same goes for statements of non-identity (except perhaps meta-linguistic versions thereof). So, for example, both "$2 = \{0,\{0\}\}$" and "$2 \neq \{0,\{0\}\}$" suffer from presupposition failure.

Aspects of Benacerraf's proposal bear a family resemblance to the solution to the Caesar problem proposed by the Scottish neo-logicists, Bob Hale and Crispin Wright (Hale and Wright 2001: Chap. 14). Their thought, roughly, is that each object is associated with a criterion of identity, indicating how that object differs from all others, or at least all others of its kind. Natural numbers, they claim, are

individuated through statements of one-to-one correspondence between concepts, while people, say, are individuated through matters of spatio-temporal continuity (or psychological continuity, or…). The difference between the neo-logicists and the Benacerraf-view articulated in this sub-section of "What Numbers Could Not Be" is that the former maintains that because of the differing criteria of identity, Caesar is *distinct* from any natural number. So the corresponding identity statement is false, not meaningless. Assuming that numbers and sets have different identity-criteria, the neo-logicist has it that "$2 = \{0,\{0\}\}$" is false and "$2 \neq \{0, \{0\}\}$" is true.

Benacerraf broaches this possibility:

> Some will want to argue that [the] identities [in question] are not senseless or unsemantical, but simply false […]. I have only the following argument to counter such a view. It will be just as hard to explain how one *knows* that they are false as it would be to explain how one knows that they are senseless, for normally we know the falsity of some identity "$x = y$" only if we know of $x$ (or $y$) that it has some characteristic that we know $y$ (or $x$) *not* to have. I know that $2 \neq 3$ because I know, for example, that 3 is odd and 2 is not, yet it seems clearly wrong to argue that we know that $3 \neq [[[0]]]$ because, say, we know that 3 has no (or seventeen, or infinitely many) members while [[[0]]] has exactly one. We know no such thing. We do not know that it does. But that does not constitute knowing that it does not. What is enticing about the view that these are all false is, of course, that they hardly seem to be open questions to which we may find the answer any day. Clearly, all the evidence is in; if no decision is possible on the basis of it, none will ever be possible.
>
> Benacerraf op. cit.: 66–67

Of course, the neo-logicist maintains that 2 does have something that $\{0,\{0\}\}$ lacks, namely its particular criterion of identity. Benacerraf concludes that "for the purposes at hand," the difference between these two views on identity "is not a very serious one," and that he "should certainly be happy with the conclusion that all identities" of the sort under study "are either senseless or false" (Benacerraf op. cit.: 67).

My proposal, to be sketched below, is perhaps closer to Benacerraf's, at least in spirit, but I take the identities in question to be *open*, and thus neither meaningless nor false. Saying the identities are meaningless precludes deciding them later, unless the meanings of our words change. Saying that the identities are false also seems to preclude deciding them later, unless we discovered we were wrong about the facts.

## 9.2 Sub-section B: "Explication and Reduction"

In this brief sub-section, Benacerraf takes up "two activities closely related to that of stating that numbers *are* sets" (Benacerraf op. cit.: 67). The first, *explication*, occurs when a theorist—mathematician or philosopher—somehow "identifies" one kind of mathematical object with another. It may be part of an explication that, for example, the natural number 3 is identical with the corresponding von Neumann ordinal. Another common explication identifies each real number with a certain Dedekind cut, and a third identifies each real number with a certain equivalence class of Cauchy sequences of rational numbers. According to Benacerraf, this sort

of "identification" does not constitute the foregoing mistake of thinking that this number just *is* this set:

> I certainly do not wish what I am arguing in this paper to militate against identifying 3 with anything you like. The difference lies in that, normally, one who identifies 3 with some particular set does so for the purpose of presenting some theory and does not claim that he has *discovered* which object 3 really is. We might want to know whether some set (and relations and so forth) would do as number surrogates. In investigating this it would be entirely legitimate to state that making such an identification, we can do with that set (and those relations) what we now do with the numbers. […] [I]t is clear that someone who says this would not claim that […] numbers were really sets all along.
>
> Benacerraf op. cit.: 68

The closely related activity of *reduction* occurs when one theory is interpreted in another. Benacerraf reports that "Gaisi Takeuti [(Takeuti 1954)] has shown that the Gödel-von Neumann-Bernays set theory is in a strong sense *reducible* to the theory of ordinal numbers less than the first inaccessible number" (Benacerraf loc. cit.). As Benacerraf put it, with a whiff of sarcasm: "No wonder numbers are sets; sets are really (ordinal) numbers, after all. *But now, which is really which?*" (Benacerraf loc. cit.). The conclusions are the same as with explication, namely that this activity does not constitute a discovery or thesis that certain objects have been identical all along.

Consider the following passage from a chapter entitled "Are sets all there is?" in Moschovakis (1994):

> [Consider] the "identification" of the […] geometric line […] with the set […] of real numbers, via the correspondence which "identifies" each point […] with its coordinate. […] What is the precise meaning of this "identification"? *Certainly not that points are real numbers*. Men have always had direct geometric intuitions about points which have nothing to do with their coordinates. […] What we mean by the "identification" […] is that the correspondence […] gives a **faithful representation** of [the line] in [the real numbers] which allows us to give arithmetic definitions for all the useful geometric notions and to study the mathematical properties of [the line] **as if points were real numbers**. […] [W]e […] discover within the universe of sets *faithful representations* of all the mathematical objects we need, and we will study set theory […] **as if all mathematical objects were sets**. The delicate problem in specific cases is to formulate precisely the correct definition of "faithful representation" and to prove that one such exists.
>
> Moschovakis (1994: 33–34)

On Penelope Maddy's gloss, "the job of set-theoretic foundations is to isolate the mathematically relevant features of a mathematical object and to find a set-theoretic surrogate with those features" (Maddy 1997: 26). Of course, surrogates, by definition, are not the real thing. Something like a replacement, supposedly one that is good for certain purposes. Benacerraf's notions of explication and reduction are useful descriptions of what is described.

The activities of explication and reduction are not only done for philosophical purposes, such as ontological economy. Maddy highlights foundational benefits:

> [T]he set-theoretic axioms […] have consequences for existing fields. […] [The] single, unified arena for mathematics provides a court of final appeal for questions of mathematical existence and proof: if you want to know if there is a mathematical object of a certain sort, you ask (ultimately) if there is a set-theoretic surrogate of that sort; if you want to know if a given statement is provable or disprovable, you mean (ultimately), from the axioms of the theory of sets.
>
> Maddy loc. cit.

> […] [V]ague structures are made more precise, old theorems are given new proofs and unified with other theorems that previously seemed distinct, similar hypotheses are traced at the basis of disparate mathematical fields, existence questions are given explicit meaning, unprovable conjectures can be identified, new hypotheses can settle old open questions, and so on.
>
> Maddy op. cit.: 34

The benefits of "explication" and "reduction," within mainstream mathematics, are, I think, even more substantial than these foundational ones. As Georg Kreisel once put it:

> […] [V]ery often the mathematical properties of a domain $D$ become only graspable when one embeds $D$ in a larger domain $D'$. Examples: (1) $D$ integers, $D'$ complex plane; use of analytic number theory. (2) $D$ integers, $D'$ $p$-adic numbers; use of $p$-adic analysis. (3) $D$ surface of a sphere, $D'$ 3-dimensional space; use of 3-dimensional geometry. Non-standard analysis [also applies] here.
>
> Kreisel (1967: 166)

To take a simple example, when one realizes that the complex plane (or, for that matter, the set-theoretic hierarchy) contains an "isomorphic copy" of the natural numbers, then one can use complex analysis (or set theory) to shed light on the natural numbers. The recent, spectacular resolution of Fermat's last theorem is a case in point, solving a problem in basic arithmetic through elliptical functions—whether or not it should turn out that the theorem has a more elementary proof.

Sometimes the identifications, or embeddings, are amazingly fruitful, shedding much light on a mathematical structure. Hermann Weyl once wrote that Riemann's approach to complex analysis should be seen

> [N]ot merely a device for visualizing the many-valuedness of analytic functions, but rather an indispensable essential component of the theory; not a supplement, more or less artificially distilled from the functions, but their native land, the only soil in which the functions grow and thrive.
>
> Weyl (1955: VII)

(See also Wilson 1992). Jamie Tappenden suggests that Weyl (1955) may not have meant to endorse this sentiment unequivocally; he was referring to his earlier, youthful exuberance (Tappenden 2005: 191n16). Nevertheless, in cases like these, I suggest that one can be forgiven for thinking that one *has* discovered the true nature of the entities in question, and that, to paraphrase Benacerraf, complex numbers have been points all along.

Nevertheless, Benacerraf says that explication and reduction are "neutral" on the metaphysical issues broached in "What Numbers Could Not Be." He notes, however, that these activities do "cast some sobering light on what it is to be an individual number." Both he and Maddy speak of "surrogates" for various mathematical objects. The underlying idea is grist for the structuralist mill. Why is it that we can get so many valuable results about a given kind of mathematical entity by studying *surrogates* for those entities? The answer, I suggest, is that in mathematics, structure is all that matters. The surrogates exemplify the structure in question and, of course, isomorphic systems are equivalent, at least in the language of the structure. So anything, in the language of the original theory, that can be said of the surrogates holds of the original structure. So, for example, a theorem of complex analysis that only refers to the natural-numbers-of-the-complex-plane is true *of* the natural numbers.

So far, so good. But what conclusions should be drawn from all of this concerning the nature of natural numbers and other mathematical entities? This takes us to the final sub-section of "What Numbers Could Not Be."

## 9.3 Sub-section C: "Conclusion: Numbers and Objects"

This subsection begins with what has become a manifesto for structuralism, as the "crux of the matter—that any recursive sequence whatever would do—suggests that what is important is not the individuality of each element but the structure which they jointly exhibit" (Benacerraf 1965: 69). Here, however, Benacerraf takes this insight in an eliminative direction:

> I therefore argue, extending the argument that led to the conclusion that numbers could not be sets, that numbers could not be objects at all; for there is no more reason to identify any individual number with any one particular object than with any other (not already known to be a number). […] Number theory is the elaboration of the properties of *all* structures of the order type of the numbers. The number words do not have single referents.
>
> Benacerraf op. cit.: 69, 70–71

This eliminative conclusion turns on presuppositions concerning what it takes to be an object and, correlatively, what it takes to be the referent of a singular term or something in the range of a first order variable. And I take *that* to be the crux of the matter, at least here, in the contrast between eliminative and *ante rem* structuralism.

Advocates of both kinds of structuralism—eliminative and *ante rem*—can agree with Benacerraf "that number words are not names of special nonnumerical entities, like sets, tomatoes, or Gila monsters" (Benacerraf op. cit.: 71). The *ante rem* structuralist holds that number words, like numerals, *are* names; but they are not names of any *non-numerical* entities. Number words are names of numbers. Numbers, in turn, are places in the natural number structure. From the *ante rem* perspective, this structure exists, and it is legitimate to have variables ranging over its places, and to have singular terms denoting individual places.

Consider Benacerraf's argument, cited just above, that "numbers could not be objects at all; for there is no more reason to identify any individual number with any one particular object than with any other (not already known to be a number)." The idea seems to be that it is not legitimate to say that numbers are objects unless we can identify each number with some object "not already known to be a number." But there seems to be no analogous requirement on any other sorts of objects. For example, no one would balk at the idea that, say, golf balls or statues, are objects, just because we have no way to identify each golf ball or each statue with an object not already known to be a golf ball or a statue.

Of course, there is a much-discussed metaphysical issue whether a given statue, say, can be identified with the material that makes it up, say a lump of clay. The statue and the clay seem to have different modal profiles: one can survive being squashed to a pancake and the other cannot. And, in philosophy of mind and in ethics there are raging debates over whether a person is to be identified with a particular body, and in philosophy of mind, whether a certain mental state, such as being in pain, is to be identified with a certain physical state. Perhaps these are analogues of the present issues concerning numbers and sets. Notice, however, that the success or failure of the identification of statues and clay lumps, persons and bodies, mental states and physical states, seems to have no bearing on the question of whether statues, persons, and mental states are objects. Prima facie, we do not need to find a non-statue-type of object to identify with the statue, or a physical entity to identify with a person or a mental state, in order to classify these as objects. Similarly, we do not need to find a unique set, or a unique anything non-numeric, in order to correctly think of numerals as names and to think of numbers as objects.

I submit that the proper orientation toward objects is a natural extension of the aforementioned view sketched in Benacerraf's first sub-section, on identity. Recall that the target there was the Fregean thesis that the range of the identity relation consists of any and all objects: "It made sense for Frege to ask of *any* two names (or descriptions) whether they named the same object or different ones" (Benacerraf op. cit.: 64). Recall that Benacerraf rejected this. He proposed instead that statements of identity carry a certain presupposition. He simply denies "that all identities are meaningful." To repeat the key passage:

> If an expression of the form "$x = y$" is to have a sense, it can be only in contexts where it clear that both $x$ and $y$ are of some kind or category $C$, and that it is the conditions which individuate things *as the same* C which are operative and determine its truth-value. […] To put the point differently, questions of identity contain the presupposition that the "entities" inquired about both belong to some general category. This presupposition is normally carried by the context or theory […]. To say that they are both "entities" is to make no presupposition at all—for everything purports to be at least that.
>
> […] There are really two correlative ways of looking at the problem. One might conclude that identity is systematically ambiguous, or else one might agree with Frege, that identity is unambiguous, always meaning sameness of object, but that (contra-Frege now) the notion of an *object* varies from theory to theory, category to category […]. This last is what I am urging.
>
> Benacerraf op. cit.: 64, 65–66, 66

I hereby urge something in the ballpark of "this last," not only as an account of identity statements, but also as an account of objecthood. If the very notion of *object* varies from theory to theory, category to category, as Benacerraf suggests, then there is no reason to expect an identification of numbers, golf balls, statues, and persons with objects "not already recognized" to be numbers, golf balls, statues, and persons, respectively. For the same reason, there is no reason to deny the honorific title of "object" to numbers.

The present proposal or, better, the present proposal sketch, fits smoothly with a slogan coined by Quine: to be an object is to be the value of a bound variable in a true theory. The basic principles of arithmetic are true, or at least their truth is not in question here (putting fictionalism and the like aside). The *ante rem* structuralist interprets arithmetic as being about a particular structure, with the first order variables ranging over the places of this structure. So, numbers—places in the natural number structure—are objects. To make the relativity explicit, the natural numbers are the objects *of arithmetic*.

When put in these Quinean terms, one might think that the issue at hand is bound up with that of so-called "absolute generality," the question of whether it is coherent to speak of "all objects whatsoever." Frege presupposed that it is indeed coherent to invoke absolutely unrestricted quantification, and that metaphysical reality, so to speak, imposes a single, determinate identity relation on this absolutely general range. It is this last that is rejected here.

Rejecting absolute generality does not resolve the problem, however. Another Quinean slogan is "no entity without identity." Perhaps there is a true theory whose objects are the natural numbers—whatever those may be—together with some other objects, say sets. By Quinean lights, *that* theory would have to have an identity relation on its combined ontology. So the combined theory would have to decide identities between numbers and the other objects, the sets.

Our question would then be whether the natural numbers of the combined theory are, as a matter of metaphysical reality, the very same objects as those in the range of the variables of arithmetic, a sort of stand-alone theory. The view proposed here maintains that there is no fact of the matter concerning *that* question. It has a kind of indeterminacy.

To put this is a larger perspective, let me sketch a philosophy of language, due to Friedrich Waismann, that I have invoked in various contexts in recent years. Then we'll briefly apply it to the matters at hand.

## 9.4 Open-Texture

The perspective begins with a thorough rejection of what Mark Wilson calls the "classical picture" of language, the thesis that the concepts we deploy are precisely delimited in all possible situations, well beyond the normal use of the words (Wilson 2006). The opposing claim, or one opposing claim, is that there is genuine indeterminacy concerning at least some possible applications of words and what

they stand for. Recall Benacerraf's claim that identity statements have a presupposition concerning theory or category. The idea here is that a similar, but less articulate, presupposition applies to just about all of language.

Waismann introduced his notion of open-texture in an attack on crude phenomenalism, the view that one can understand any cognitively significant statement in terms of sense-data. The failure of the verificationist program:

> is not, as has been suggested, due to the poverty of our language which lacks the vocabulary for describing all the minute details of sense experience, nor is it due to the difficulties inherent in producing an *infinite* combination of sense-datum statements, though all these things may contribute to it. In the main it is due to a factor which, though it is very important and really quite obvious, has to my knowledge never been noticed—to the 'open-texture' of most of our empirical concepts.
>
> Waismann [1945] (1968: 121)

Here is one of the thought experiments Waismann uses to illustrate open-texture:

> Suppose I have to verify a statement such as 'There is a cat next door'; suppose I go over to the next room, open the door, look into it and actually see a cat. Is this enough to prove my statement? […] What […] should I say when that creature later on grew to a gigantic size? Or if it showed some queer behavior usually not to be found with cats, say, if, under certain conditions it could be revived from death whereas normal cats could not? Shall I, in such a case, say that a new species has come into being? Or that it was a cat with extraordinary properties? […] The fact that in may cases there is no such thing as a conclusive verification is connected to the fact that most of our empirical concepts are not delimited in all possible directions.
>
> Waismann op. cit.: 121–122

The last cited "fact" is the key insight behind this philosophy of language. Language users introduce terms to apply to certain objects or kinds of objects, and, of course, the terms are supposed to fail to apply to certain objects or kinds of objects. The point, after all, is to make distinctions. However, as we introduce terms, and use them in practice, we cannot be sure that every possible situation is covered, one way or the other. This applies even in science. Indeed, Waismann applies the concept of open-texture to a concept that is now sometimes invoked as a paradigm natural kind:

> The notion of gold seems to be defined with absolute precision, say by the spectrum of gold with its characteristic lines. Now what would you say if a substance was discovered that looked like gold, satisfied all the chemical tests for gold, whilst it emitted a new sort of radiation? 'But such things do not happen.' Quite so; but they *might* happen, and that is enough to show that we can never exclude altogether the possibility of some unforseen situation arising in which we shall have to modify our definition. Try as we may, no concept is limited in such a way that there is no room for any doubt. We introduce a concept and limit it in *some* directions; for instance we define gold in contrast to some other metals such as alloys. This suffices for our present needs, and we do not probe any farther. We tend to *overlook* the fact that there are always other directions in which the concept has not been defined. […] We could easily imagine conditions which would necessitate new limitations. In short, it is not possible to define a concept like gold with absolute precision; i.e., in such a way that every nook and cranny is blocked against entry of doubt. That is what is meant by the open-texture of a concept.
>
> Waismann op. cit.: 122–123

Waismann concludes: "Every description stretches, as it were, into a horizon of open possibilities: However far I go, I shall always carry this horizon with me" (Waismann op. cit.: 124).

Let $R$ be an $n$-place predicate from natural language. Say that $R$ exhibits *open-texture* if there could be objects $p_1, \ldots, p_n$, such that nothing in the established use of $R$, or the non-linguistic facts, determines that $R$ holds of $p_1, \ldots, p_n$, and nothing in the established use of $R$, nor the non-linguistic facts, determines that $R$ fails to hold of $p_1, \ldots, p_n$. In effect, the sentence $Rp_1 \ldots p_n$ is left open by the use of the language, to date. The claim here is that the identity relation is subject to open-texture.

The phrase "open-texture" does not appear in Waismann's treatment of the analytic-synthetic distinction in a lengthy article published serially in *Analysis* (Waismann 1949, 1950, 1951a, b, 1952, 1953), but the notion clearly plays a central role there. He observes that language is an evolving phenomenon. As new situations are encountered, and as new scientific theories develop, the extensions of predicates change. Sometimes the predicates become sharper, which is what someone who accepts open-texture would predict. As new cases are encountered, the predicate in question is extended to cover them, one way or the other. When things like this happen, there is often no need to decide—and no point in deciding—whether the application of a given predicate to a novel case represents a change in its meaning or a discovery concerning the term's old meaning, going on as before, as Wittgenstein might put it.

Waismann said, in an early article in the series: "there are no *precise rules* governing the use of words like "time," "pain," etc., and that consequently to speak of the 'meaning' of a word, and to ask whether it has, or has not changed in meaning, is to operate with too blurred an expression" (Waismann 1951a: 53). The key word here is "precise." I take it that the "blurred expression" we are trying to "operate with" is something like "means that same as." There is genuine indeterminacy concerning that relation.

The opposing thesis, that there is always one and only one correct way to go on consistent with the meaning of the term is what Wilson calls the classical picture. This badly misrepresents the nature of language:

> Simply […] to refer to "the" ordinary use [of a term] is naive. […] [The] whole picture is in a state of flux. One must indeed be blind not to see that there is something unsettled about language; that it is a living and growing thing, adapting itself to new sorts of situations, groping for new sorts of expression, forever changing.
>
> Waismann (1951b: 122–123)

Toward the end of the series, Waismann writes: "What lies at the root of this is something of great significance, the fact, namely, that language is never complete for the expression of all ideas, on the contrary, that it has an essential *openness*" (Waismann 1953: 81–82).

Waismann's official definition of open-texture in Waismann [1945] (1968) limits it to the application of empirical predicates (or their negations) to hitherto unconsidered, and even undreamt of, cases. In Shapiro (2006), I argue that the mathematical notion of a computable function (of natural numbers) is, or at least was, circa 1936, subject to open-texture. The notion of a recursive (or Turing computable) function served to sharpen it, to the extent that, now, Church's thesis is more or less uncontroversially true. Nevertheless, to ask whether the notion of a recursive (or Turing computable) function corresponds exactly to *the* notion of a computable function—the one implicit in the earlier mathematical treatments—is to operate with too blurred an expression. Again, the blurred expression is something like "means the same as." The mathematical work served to sharpen or replace the pre-theoretic, intuitive notion of computability.

Again, the proposal here is that the very notion of identity, at least within mathematics, is similarly subject to open-texture. As new objects—and new structures—are encountered and incorporated into theories, new identity statements emerge, statements that, at least at first, are genuinely indeterminate.

Here is a sort of rational reconstruction of the current situation, amply described in *Benacerraf 1965*, but recast in terms of open-texture. I am not claiming that this is any more than a cartoon sketch of how something like the present state of play could have come about.

At first, a community of mathematicians developed a theory of arithmetic. They have a pure theory, which concerns only the natural number structure. Say it is ordinary second order Peano-Dedekind arithmetic (to make sure that the theory is categorical, and thus concerns only one structure). The mathematicians also have an applied theory, showing how numbers—places in the structure—can be used in various ways, such as determining the cardinalities of various collections, balancing their checkbooks, and the like.

A bit later, the community develops a richer mathematical theory, say set theory, and think they can use the resources of the richer theory to shed light on the natural number structure. So they formulate a combined theory, one that has both variables ranging over natural numbers and variables ranging over sets. They might have two different variable-sorts, or else they may invoke a single sort, covering numbers and sets, and introduce a predicate for natural numbers and a predicate for sets. This difference does not matter here.

So now we confront a version of the question posed at the end of the previous section. Is it the case that the "natural numbers" of the combined theory are the very same objects as those studied in the pure theory of arithmetic? In terms of *ante rem* structuralism, the question is whether the range of the arithmetic variables in the combined theory (or the extension of the "natural number" predicate in that theory) consists of the places in *the* natural number structure, the subject matter of the pure theory of arithmetic. The view here is that there is no fact of the matter concerning *that* question. It has a kind of indeterminacy. To ask if the original natural numbers are the same as the new natural numbers, is the same sort of question as asking if the arithmetic terminology in the original theory has the same meaning as its counterpart in the new theory. And that, to paraphrase Waismann, is to operate with

too blurred an expression. As above, the blurred expression here is something like "means the same as" or, in more ontological mode, "has the same subject matter." There is genuine indeterminacy concerning that relation, vagueness if you will.

So let us turn to the combined theory, putting the "past" pure theory aside for the time being. In the new theory, the community of mathematicians might need, or at least want, an identity relation. If they take the Quinean slogan, no entity without identity, seriously, then they will *need* an identity relation, in order to satisfy themselves that they are dealing with entities. Slogans aside, the community might *want* an identity relation, for matters of convenience, ontological economy, or what have you.

In other words, the mathematical/linguistic community in question might need to or want to decide identity statements between the natural numbers of the new theory and the sets of the new theory. Satisfying this want or need is not a matter of discovering some fact about the world. Nothing that they have said or done to date determines whether a given natural number is the same as or different from a given set. Such is open-texture. The situation is just like that in Waismann's thought experiments, where it is noted that nothing we have done to date determines whether the envisioned creature (if that is what it is) is a cat, or whether the envisioned substance is gold.

Recall Waismann's quip: "Every description stretches, as it were, into a horizon of open possibilities: However far I go, I shall always carry this horizon with me" (Waismann [1945] 1968: 124). The present claim is that, at least in mathematics, the "horizon" applies to the identity relation as much as to any other notion. As new structures are defined and studied, and as these new structures are related to the existing ones, identity issues can arise, and can be settled.

In the rational reconstruction at hand, the mathematical-cum-linguistic community has at least two choices. One is to identity the natural numbers with certain sets, such as the finite von Neumann ordinals. If they choose, they can make the identification permanent. So as a matter of linguistic fiat, the numeral "2" would denote the set $\{\phi,\{\phi\}\}$. The community would thereby sharpen the identity relation, similar to what happens when the use of any open-texture term is extended to cover new cases, one way or the other. The resulting theory will extend the old, pure arithmetic. No theorems would be lost. The reason this would go so smoothly is that the finite von Neumann ordinals form an ω-sequence, and so anything, in the language of arithmetic, that is true of this set is also true of the original natural numbers.

A second option for the mathematicians is to insist that natural numbers are distinct from sets. That is, if $m$ is a natural number and $s$ is a set, then $m \neq s$. They can still, to use Benacerraf's term, *explicate* the natural numbers as certain sets (say the finite von Neumann ordinals), for certain purposes, or they can reduce arithmetic to set theory in various ways. However, because it is an explication, or reduction, the community would not be saying that natural numbers *are* sets. The community might even find it convenient to use different explications at different times, or even at the same time. Some might prove convenient for some purposes, others for other purposes.

From the perspective urged here, there is no question, no matter of fact, concerning whether one or the other of these options is the correct way to "go on as before," as Wittgenstein might have put it. Again, to think that one or the other option—a permanent identification or a range of explications—is *the* correct option is to fall into the classical picture of language, the thesis that the concepts we deploy are precisely delimited in all possible situations. This applies especially to the identity relation.

Concerning the actual world of mathematics, Wilson once noted that "any notion that the reals should not be identified with sets represents as great a misunderstanding of mathematical ontology as the claim that they should" (Wilson 1981). That orientation is sanctioned here, through the notion of open-texture.

To reiterate, the foregoing rational reconstruction, and the ensuing observations, are meant to support the conclusion of the previous section and, for that matter, that of Shapiro (1997), namely that natural numbers are places in the natural number structure and that, as such, they are fully bona fide objects. They are as bona fide as any other objects, be they avocados, statues, or golf balls.

## References

Burgess, J. P., & Rosen, G. (1997). *A subject with no object: Strategies for nominalistic interpretation of mathematics*. Oxford: Oxford UP.

Dummett, M. A. E. (1991). *Frege: Philosophy of mathematics*. Cambridge, MA: Harvard UP.

Frege, G. [1884] (1960). *Die Grundlagen der Arithmetik, Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Breslau: Verlag von Wilhelm Koebner [The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number (2nd Ed., J. L. Austin, Trans.). New York: Harper].

Geach, P. (1967). Identity. *The review of metaphysics, 21*(1), 3–12.

Hale, B., & Wright, C. (2001). *The reason's proper study*. Oxford: Oxford UP.

Hellman, G. (1989). *Mathematics without numbers*. Oxford: Oxford UP.

Kreisel, G. (1967). Informal rigour and completeness proofs. In I. Lakatos (Ed.), *Problems in the philosophy of mathematics* (pp. 138–186). Amsterdam: North Holland Publishing Company.

Maddy, P. (1997). *Naturalism in mathematics*. Oxford: Oxford UP.

Moschovakis, Y. (1994). *Notes on set theory*. New York: Springer.

Parsons, C. (1990). The structuralist view of mathematical objects. *Synthese, 84*(3), 303–346.

Shapiro, S. (1997). *Philosophy of mathematics: Structure and ontology*. Oxford: Oxford UP.

Shapiro, S. (2006). Computability, proof and open-texture. In R. Janusz, A. Olszewski, & J. Wokeński (Eds.), *Church's thesis after 70 years* (pp. 420–455). Frankfurt: Ontos Verlag.

Takeuti, G. (1954). Construction of the set theory from the theory of ordinal numbers. *Journal of the Mathematical Society of Japan, 6*, 196–220.

Tappenden, J. (2005). Proof style and understanding in mathematics I: visualization, unification and axiom choice. In P. Mancosu, K. F. Jørgensen, & S. A. Pederson (Eds.), *Visualization, explanation and reasoning styles in mathematics* (pp. 147–207). Dordrecht: Springer.

Waismann, F. (1949). Analytic-synthetic I. *Analysis, 10*(2), 25–40.

Waismann, F. (1950). Analytic-synthetic II. *Analysis, 11*(2), 25–38.

Waismann, F. (1951a). Analytic-synthetic III. *Analysis, 11*(3), 49–61.
Waismann, F. (1951b). Analytic-synthetic IV. *Analysis, 11*(6), 115–124.
Waismann, F. (1952). Analytic-synthetic V. *Analysis, 13*(1), 1–14.
Waismann, F. (1953). Analytic-synthetic VI. *Analysis, 13*(4), 73–89.
Waismann, F. [1945] (1968). Verifiability. In A. Flew (Ed.) *Logic and language* (pp. 117–144), Oxford: Basil Blackwell.
Weyl, H. (1955). *The Concept of a Riemann Surface* (G. Maclane, English translation of the 3rd revised German edition of 1913). Reading, MA: Addison-Wesley.
Wilson, M. (1981). The double standard in ontology. *Philosophical Studies, 39*(4), 409–427.
Wilson, M. (1992). Frege: The royal road from geometry. *Noûs, 26*(2), 149–180.
Wilson, M. (2006). *Wandering significance*. Oxford: Oxford UP.

# Part III
# Supertasks

If we take the stand that "nonconstructive" procedures—i.e., procedures that require us to perform infinitely many operations in a finite time—are conceivable,* though not *physically* possible (owing mainly to the existence of a limit to the velocity with which physical operations can be performed), then we can say that there does "in principle" exist, a verification/refutation procedure for number theory.

*E.g., if one has an infinite series of operations to perform, say $S_1$, $S_2$, $S_3$,… and one is able to perform $S_1$ in 1 min, $S_2$ in 1/2 min, $S_3$ in 1/4 min, etc.; then in 2 min one will have completed the whole infinite series.

Benacerraf and Putnam ([1964] 1983: 20)

# Chapter 10
# Tasks, Subtasks and the Modern Eleatics

**Jon Pérez Laraudogoitia**

## 10.1 Introduction

In a frequently quoted paragraph, Benardete presented the "paradox of the gods" in the following terms:

> A man decides to walk one mile from A to B. A god waits in readiness to throw up a wall blocking the man's further advance when the man has travelled ½ mile. A second god (unknown to the first) waits in readiness to throw up a wall of his own blocking the man's further advance when the man has travelled ¼ mile. A third god…, etc. *ad infinitum*. It is clear that this infinite sequence of mere *intentions* (assuming the contrary-to-fact conditional that each god would succeed in executing his intention if given the opportunity) logically entails the consequence that the man will be arrested at point A; he will not be able to pass beyond it, even though not a single wall will in fact be thrown down in his path. The before-effect here will be described by the man as a strange field of force blocking his passage forward.
>
> Benardete (1964: 259–260)

In the present paper, I set myself four tasks. The first is to construct a generalization of the Benardete Dichotomy that allows us to reveal its true scope. I then defend this generalization by demonstrating that it may quite clearly be analyzed in Newtonian terms which are also applicable to the original Benardete Dichotomy. The demonstrated possibility of a physically consistent Benardete Paradox suggests the introduction of a new concept, the concept of subtask, that unifies and systematizes the study of the paradox and its many generalizations. Finally I propose a purely mechanical model of the paradox of the gods that allows me to formulate an argument against one of the most widely accepted critical analyses of the paradox available today. So my defence of the subtask of Benardete's gods has an affinity

J.P. Laraudogoitia (✉)
Department of Logic and Philosophy of Science, University of the Basque Country (UPV/EHU), Donostia, Spain
e-mail: jon.perez@ehu.eus



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_10

with Benacerraf's classic defence of the logical possibility of the supertask of Thomson's lamp (see Benacerraf 1962): in both cases, the idea is to vindicate the legitimacy of two notions which, as we shall see, are related in a peculiar way.

## 10.2 The Logical Structure of Two Generalizations of the Paradox of the Gods

In the Benardete paradox, a man (who, to all intents and purposes, may be considered a particle) is prevented from moving. Preventing the movement of a particle entails imposing a very particular form on its World line (namely, the form of a straight line). In the first generalization I shall present (let us call it G1), the gods ($god_1$, $god_2$, $god_3$, $god_4$, …) will make particle Q follow the World line $L \equiv \{(t, \mathbf{L}(t))/\ t_1 < t < t_2\}$ between the instants $t_1$ and $t_2$, where $\mathbf{L}()$ represents a continuous function of time t but which, for any other purpose, may have an arbitrarily complex form (corresponding to a highly intricate evolution). In this paper, I assume (and this is also implicit in Benardete) that the World lines of all the particles are always continuous functions of time. The meaning of the term "generalizations" that appears in the title of this section is also clear: it will be shown that Benardete's gods are not only able to prevent a particle from moving, but also that they are able to direct its evolution in time according to any continuous function of t. This means that any admissible form of the evolution of a particle (and, as we shall see at the end of this section, even of any system of particles at the most numerably infinite) is susceptible of being reproduced by following the essential features of the schema proposed by Benardete. In the initial instant $t_1$, let Q be at point $\mathbf{L}(t_1)$. As a general rule, we refer to $\mathbf{L}(t)$ as the point of space corresponding to the World line L at instant t. Although Benardete is concerned with counterfactual conditionals in the quote, his paradox may be conveniently analyzed and generalized without them (several authors have as a matter of fact done so). All one needs is to stipulate that the gods decide to follow a certain plan of action (which, we assume, it is in their power to follow) and then specify this plan in terms of material conditionals. In any event, should anyone consider this to violate the original sense of Benardete's work, one could still interpret my analyses as variations on the paradox of the gods rather than as bona fide generalizations.

### *10.2.1 The First Generalization (G1)*

In G1, I assume, in particular, that the gods decide to follow the plan specified by the following set of sentences:

$C_1^{(1)}$: if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/1 from **L**(t), then for every instant t*, with $t < t^* < t_2$, $god_1$ will ensure that, at t*, Q is at a distance $\leq$ 1/1 from **L**(t*).

$\mathbf{C}_2^{(1)}$: if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/2 from **L**(t), then for every instant t*, with $t < t^* < t_2$, $god_2$ will ensure that, at t*, Q is at a distance $\leq$ 1/2 from **L**(t*).

$C_3^{(1)}$: if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/3 from **L**(t), then for every instant t*, with $t < t^* < t_2$, $god_3$ will ensure that, at t*, Q is at a distance $\leq$ 1/3 from **L**(t*).

.....

$C_n^{(1)}$: if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/n from **L**(t), then for every instant t*, with $t < t^* < t_2$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from **L**(t*).

.....

In symbolic notation, with "$d_t(A,B)$" meaning "distance between A and B at the instant t" and "$g_nH$" meaning "$god_n$ ensures that H," we have respectively[1]:

$$C_1^{(1)} : \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/1 \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_1 d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/1)]).$$

$$C_2^{(1)} : \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/2 \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_2 d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/2)]).$$

$$C_3^{(1)} : \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/3 \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_3 d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/3)]).$$

. . . . . .

$$C_n^{(1)} : \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/n \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_n d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/n)]).$$

. . . . . .

It is easy to see that if each $C_n^{(1)}$ (n = 1,2,3,…) is true then C: particle Q follows the World line L between instant $t_1$ and instant $t_2$. In other words C: $\forall t\ (t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) = 0)$ will also be true:

$$(\mathbf{I})(\wedge_n C_n^{(1)}) \rightarrow C.$$

*Proof*: All one needs to do is to confirm that the negation of C is incompatible with the conjunction of all the $C_n^{(1)}$. Indeed, if C is false, then at some instant t with $t_1 < t < t_2$, particle Q is at a distance > 0 from **L**(t):

$$\exists t(t_1 < t < t_2 \wedge d_t(Q, \mathbf{L}(t)) > 0).$$

Let us call one of these instants T:

$$t_1 < T < t_2 \wedge d_T(Q, \mathbf{L}(T)) = \delta > 0 \tag{10.1}$$

and let N be a positive integer such that $\delta > 1/N$. We assumed that the function **L**() is continuous, that at $t_1$ the particle was on L: $d_{t1}(Q, \mathbf{L}(t_1)) = 0$, and that the World line of Q is continuous. This implies that, between $t_1$ and T, particle Q was at all possible distances from **L**(t) between 0 and $\delta$. To put it more precisely:

$$\forall \alpha(0 < \alpha < \delta \rightarrow \exists t(t_1 < t < T \wedge d_t(Q, \mathbf{L}(t)) = \alpha)).$$

As $0 < 1 / N < \delta$,

$$\exists t(t_1 < t < T \wedge d_t(Q, \mathbf{L}(t)) = 1/N).$$

Let us call one of these instants $T_N$. At $T_N$, with $t_1 < T_N < T$, Q must have been at a distance 1/N from $\mathbf{L}(T_N)$:

$$t_1 < T_N < T \wedge d_{TN}(Q, \mathbf{L}(T_N)) = 1/N. \tag{10.2}$$

However, from (10.1) and (10.2) it follows that

$$t_1 < T_N < t_2 \wedge d_{TN}(Q, \mathbf{L}(T_N)) = 1/N \tag{10.3}$$

and, particularizing $C_N^{(1)}$ to the case $t = T_N$,

$$\begin{aligned} t_1 < T_N < t_2 \rightarrow [d_{TN}(Q, \mathbf{L}(T_N)) = 1/N \rightarrow \forall t^*(T_N < t^* < \\ t_2 \rightarrow g_N d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/N)]. \end{aligned} \tag{10.4}$$

From (10.3) and (10.4):

$$\forall t^*(T_N < t^* < t_2 \rightarrow g_N d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/N). \tag{10.5}$$

But from (10.1) and (10.2) we also know that $T_N < T < t_2$ so that, particularizing (10.5) to the case $t^* = T$, it follows that

$$g_N d_T(Q, \mathbf{L}(T)) \leq 1/N. \tag{10.6}$$

In other words, given that $g_n H$ implies H (according to the formalization noted above):

$$d_T(Q, \mathbf{L}(T)) \leq 1/N). \tag{10.7}$$

However, from (10.1),

$$d_T(Q, \mathbf{L}(T)) = \delta \tag{10.8}$$

which contradicts (10.7) by virtue of the fact that, as I said at the beginning, $\delta > 1/N$. The contradiction proves that the negation of C is incompatible with the conjunction of all the $C_n^{(1)}$. Therefore, when all the $C_n^{(1)}$ are true, C is too. ┤

We may proceed further because if C is true, then every $C_n^{(1)}$ is also true (owing to the fact that the main conditional occurring in $C_n^{(1)}$ would then be vacuously true) and, therefore, so would be its conjunction:

$$(\mathbf{II})\ C \rightarrow (\wedge_n C_n^{(1)}).$$

*Proof*: From the properties of the material conditional of propositional logic we know that, for each positive integer n:

$$\forall t(d_t(Q, \mathbf{L}(t)) = 0 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/n \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_n d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/n)])$$

and then

$$\forall t\ ([t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) = 0] \rightarrow [t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/n \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_n d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/n)]]).$$

From this and C it is clear that, for each positive integer n:

$$\forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/n \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_n d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/n)]).$$

In other words, every $C_n^{(1)}$ is also true. ┤

We then conclude that C is equivalent to the infinite conjunction of the $C_n^{(1)}$:

$$(\mathbf{III})\ C \leftrightarrow (\wedge_n C_n^{(1)}).$$

*Proof*: From (**I**) and (**II**). ┤

### *10.2.2 The Interesting Generalization (G2)*

To generalize the paradox of the gods suitably, a stipulation is required that prevents any god from being able to act on its own on Q (even in a way compatible with the $C_n^{(1)}$) beyond what is explicitly permitted by $C_n^{(1)}$. We thus construct the

second generalization G2 so that it is identical to G1 except for the fact that each $god_n$ fulfils $C_n^{(2)}$ rather than $C_n^{(1)}$.

$C_n^{(2)}$: if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/n from $\mathbf{L}(t)$, then for every instant t*, with $t < t^* < t_2$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from $\mathbf{L}(t^*)$. If at some instant t, with $t_1 < t < t_2$, Q does not come to be at a distance 1/n from $\mathbf{L}(t)$, then $god_n$ shall not apply any force at all on Q at t.

That is, with "$\neg g_n(Q, t)$" meaning "$god_n$ does not apply any force at all on Q at t":

$$C_n^{(2)} : \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) = 1/n \rightarrow \forall t^*(t < t^* < t_2 \rightarrow g_n d_{t^*}(Q, \mathbf{L}(t^*)) \leq 1/n)]) \wedge \forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) \neq 1/n \rightarrow \neg g_n(Q, t)]).$$

Finally, what makes G2 specifically interesting is that:

(**IV**) In generalization G2, no $god_n$ exerts individually any force at all on Q.

*Proof*: Since $(\wedge_n C_n^{(2)}) \rightarrow C$ (given that $(\wedge_n C_n^{(2)}) \rightarrow (\wedge_n C_n^{(1)})$ and that $(\wedge_n C_n^{(1)}) \rightarrow C$), it follows that, in generalization G2, particle Q evolves according to $\mathbf{L}(t)$, i.e., Q follows the World line L between instant $t_1$ and instant $t_2$ and, therefore, the force acting on it is $m\, d^2\mathbf{L}/dt^2$.

Since, by C: $\forall t\ (t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) = 0)$, it follows that

$$\forall t(t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) \neq 1/n). \tag{10.9}$$

However, the second term of the conjunction $C_n^{(2)}$ implies:

$$\forall t\ (t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) \neq 1/n) \rightarrow \forall t(t_1 < t < t_2 \rightarrow \neg g_n(Q, t)). \tag{10.10}$$

From (10.9) and (10.10): $\forall t\ (t_1 < t < t_2 \rightarrow \neg g_n(Q, t))$ and, therefore, $god_n$ does not apply any force at all on Q. ⊣

From (**IV**) it follows that, if $E_i$ is the sentence:

$E_i$ : $god_i$ applies the force $m\, d^2\mathbf{L}/dt^2$ on Q, between instant $t_1$ and instant $t_2$, by pushing it by contact himself

then $E_i \wedge (\wedge_n C_n^{(2)})$ is logically inconsistent.

### 10.2.3 The Interesting Generalization: Many Particles

Instead of directing the evolution of a single particle Q without individually exerting any force at all on it, a numerable infinity of gods may likewise direct the evolution of any material body composed of particles (even of a numerable infinity of particles) without applying forces individually on any of them, which leads to a new generalization of the Benardete paradox.

Let $\Omega$ be a material body composed of the numerable infinity of particles $Q_1$, $Q_2$, $Q_3$, $Q_4$, … and let us suppose that the gods want $(\forall i)$ $Q_i$ to follow the World line $L_i \equiv \{(t, \mathbf{L}_i(t))/\ t_1 < t < t_2\}$ between $t_1$ and $t_2$. If we use $\pi(n)$ to represent the n-th prime number ($\pi(1) = 2$, $\pi(2) = 3$, $\pi(3) = 5$, etc.) and $\pi(m,n)$ to represent the m-th power of $\pi(n)$ ($\pi(1,1) = 2$, $\pi(2,1) = 2^2$, $\pi(3,2) = 3^3$, etc.), then there is enough $(\forall i)$ for the numerable infinity of gods: $god_{\pi(1,i)}$, $god_{\pi(2,i)}$, $god_{\pi(3,i)}$, $god_{\pi(4,i)}$, … to direct the evolution of $Q_i$ along the World line $L_i$ between $t_1$ and $t_2$, following the procedure seen in the previous generalization G2. What we have done is simply to select a certain infinite numerable set S of gods and perform in it a partition into a numerable infinity of classes $S_i$, each being formed by a numerable infinity of gods. Then the gods of $S_i$ direct the evolution of $Q_i$ exactly as in G2.

## 10.3 The Physics Underlying the Paradox (I)

Let us now recover G2 in order to analyze in a sufficiently general way the conceptual difficulties to which the paradox of the gods leads from a physical perspective. If no god exerts any force on Q, it is because Q is already following the World line L. Fulfilling the conditions $C_n^{(2)}$ implies that the force $\mathbf{F} = m\ d^2\mathbf{L}/dt^2$ is applied on Q and takes it along the World line L. The problem of explaining who exerts that force on Q is undoubtedly a genuine problem, but it is a physical problem, not a logical problem. As we already saw, Benardete claims that "The before-effect here will be described by the man as a strange field of force blocking his passage forward" (Benardete 1964: 260). What explanation might there be for the force $\mathbf{F} = m\ d^2\mathbf{L}/dt^2$ that the gods exert on Q, without introducing any "strange field of force"?

### 10.3.1 A More Detailed Specification of G2

In the first place, I shall suppose that if any god interacts with Q, it does so only by contact, so that when in $C_n^{(2)}$ it states:

> (A) if, at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/n from $\mathbf{L}(t)$, then for every instant t*, with $t < t^* < t_2$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from $\mathbf{L}(t^*)$.

this means that if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/n from **L**(t), then for every instant t*, with $t < t^* < t_2$, $god_n$ shall exert by contact on Q the force necessary to ensure that, at t*, Q is at a distance $\leq$ 1/n from **L**(t*).

Specifying further, I shall assume that for every instant t*, with $t < t^* < t_2$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from **L**(t*):

(a) by positioning itself at instant t at a distance 1/n from **L**(t) with Q between it and **L**(t) (aligned therefore with Q and **L**(t))[2] and
(b) for every instant t′ with $t < t' < t^*$, by positioning itself at a distance 1/n from **L**(t′) aligned with Q and **L**(t′).

This guarantees that $god_n$ will exert by contact on Q the force necessary to ensure that Q, at t*, is at a distance $\leq$ 1/n from **L**(t*).

Consequently, (A) implies that, if at some instant t, with $t_1 < t < t_2$, Q comes to be at a distance 1/n from **L**(t), then $god_n$ positions itself, at t, at a distance 1/n from **L**(t) with Q between it and **L**(t) (aligned therefore with Q and **L**(t)), and for every instant t*, with $t < t^* < t_2$, $god_n$ positions itself at a distance 1/n from **L**(t*) aligned with Q and **L**(t*).

In symbolic notation, with "$G_nQ_{tt2}$" meaning "$god_n$ positions itself, at t, at a distance 1/n from **L**(t) with Q between it and **L**(t) (aligned therefore with Q and **L**(t)), and for every instant t*, with $t < t^* < t_2$, $god_n$ positions itself at a distance 1/n from **L**(t*) aligned with Q and **L**(t*)," we have:

$$\forall t[(t_1 < t < t_2 \wedge d_t(Q, \mathbf{L}(t)) = 1/n) \rightarrow G_nQ_{tt2}]. \tag{10.11}$$

### *10.3.2 Excursus: Interaction by Contact and Impenetrability*

The key to my reasoning is to be found from here on in a careful consideration of what interaction by contact amounts to in classical mechanics. The first thing one notices is how unfortunate the expression "interaction by contact" is. (Despite this, I use it freely in the present paper, following common practice). Since mere contact lacks the causal powers associated with the idea of interaction, it would be better to speak of "interaction in contact conditions." I shall be looking here at a special type of interaction in contact conditions; not, e.g., through gravitational interaction in contact conditions (which in general lacks any special theoretical interest), but rather interaction in contact conditions based on the impenetrability of the bodies involved. Impenetrability is a basic postulate implicit in mechanics. (I also implicitly assume it in the present context). Its principal virtue (besides its smooth empirical fit) lies in the fact that it avoids complicating the theory with more complex hypotheses about the internal structure of matter. The reason for my interest in interaction in contact conditions based on impenetrability is that what is usually called interaction by contact amounts to just this, i.e. to an interaction in

contact conditions based on impenetrability. But, as we shall see in this section, the truly central point is the idea of interaction based on impenetrability (whether in contact conditions or not), which thus becomes a more general concept than the classic one of interaction by contact and one that is certainly compatible with it.

For the sake of simplicity, in what follows I consider that:

(a) all the relevant material bodies have volume and are rigid or are composed of rigid parts and that
(b) the only kind of interaction this provides is interaction based on impenetrability.

We must now specify what interaction based on impenetrability is exactly. I therefore use the complex term "interaction by impenetrability" and introduce the principle of interaction by impenetrability as follows:

PII: two material bodies interact by impenetrability at t if and only if, should they not interact at t (everything else remaining the same),[3] a massive part of one of them would interpenetrate with a massive part of the other at instants of time $t' > t$ arbitrarily close to t.

By way of illustration, let us consider a cylinder C (of any finite radius) at rest on the X axis such that its own axis occupies the interval $x = 0$ to $x = 1$. Particle Q approaches it at a certain constant velocity from the region $x < 0$. By PII, Q interacts by impenetrability with C at instant t when it arrives at $x = 0$, because if it did not interact, it would pass the point in question and would interpenetrate with C at instants of time $t' > t$ arbitrarily close to t (in other words, "a massive part of Q would interpenetrate with a massive part of C at instants of time $t' > t$ arbitrarily close to t").

Once the existence of such an interaction has been established, the relevant laws of conservation (i.e., in the case at hand, the laws of conservation of momentum and kinetic energy) would be applied in the conventional way together with the implicit postulate of impenetrability (mentioned above, and not to be confused with PII), in order to predict the future state of the system formed by Q and C. This example suffices to justify the idea that, as I said above, the usual textbook notion of interaction by contact is a form of interaction by impenetrability. It must be clear now that PII tells us when there is interaction by impenetrability and, in case there is, that the conventional "machinery" of classical mechanics should be used to make quantitative predictions.[4]

### *10.3.3 The Dynamics Underlying G2*

We shall now apply PII to understand the origin of the force that the gods exert on Q. Let us suppose that, with no loss of generality, the World line L is sufficiently complicated so as to not have parts that depend linearly on t (if the World line of Q depends linearly on t, then Q moves freely since, in that case, $m\, d^2\mathbf{L}/dt^2 = \mathbf{F} = 0$). In

other words, let us suppose that the force that acts on a particle with the world line L is never null: $\forall t$ with $t_1 < t < t_2$ is $\mathbf{F}(t) \neq 0$.

We shall need three prior logically connected results.

(**V**) If, at some instant $t^o$ between $t_1$ and $t_2$, Q did not interact in any way with the gods, so that the force on it were $\mathbb{F}(t^o) = 0$ (everything else remaining the same, i.e. everything that is compatible with non-interaction between Q and the gods at $t^o$ remaining the same, which in particular means that the new spatial location $\mathbb{L}(t)$ of Q is $\mathbb{L}(t) = \mathbf{L}(t)$ in all instants of the interval $(t_1, t^o)$), then, for any $\varepsilon > 0$, the spatial location $\mathbb{L}(t)$ of Q cannot coincide with $\mathbf{L}(t)$ at all instants of the interval $(t^o, t^o + \varepsilon)$.

*Proof*: If $\mathbb{L}(t) = \mathbf{L}(t)$ for all $t \in (t_1, t^o + \varepsilon)$, then $\mathbb{F}(t) = m\, d^2\mathbb{L}/dt^2|_t = m\, d^2\mathbf{L}/dt^2|_t = \mathbf{F}(t) \neq 0$ is also fulfilled for all $t \in (t_1, t^o + \varepsilon)$ with $t_1 < t < t_2$, which contradicts the starting assumption that, for the instant $t^o \in (t_1, t_2)$, Q does not interact in any way with the gods ($\mathbb{F}(t^o) = 0$).[5] Therefore, whatever $\varepsilon > 0$ is, the spatial location $\mathbb{L}(t)$ of Q cannot coincide with $\mathbf{L}(t)$ at all instants of the interval $(t_1, t^o + \varepsilon)$, although it does in fact coincide (as stated above) at all instants of the interval $(t_1, t^o)$. As a consequence, for any $\varepsilon > 0$, the spatial location $\mathbb{L}(t)$ of Q cannot coincide with $\mathbf{L}(t)$ at all instants of the interval $(t^o, t^o + \varepsilon)$. In symbols:

$$\forall \varepsilon > 0 \exists \beta \ (t^o < \beta < t^o + \varepsilon) \ d_\beta(Q, \mathbf{L}(\beta)) \neq 0 \quad \dashv \tag{10.12}$$

(**VI**) If, for any $\varepsilon > 0$, the spatial location $\mathbb{L}(t)$ of Q cannot coincide with $\mathbf{L}(t)$ at all instants of the interval $(t^o, t^o + \varepsilon)$, then there is an infinite sequence of positive integers $z_i$ ($z_1 < z_2 < z_3 < \ldots$) and an infinite sequence of real numbers $t^\wedge_i$ ($t^\wedge_1 > t^\wedge_2 > t^\wedge_3 > \ldots$) convergent to $t^o$, such that, $\forall i$, $god_{zi}$ positions itself at $t^\wedge_i$ at a distance $1/z_i$ from $\mathbf{L}(t^\wedge_i)$ with Q between it and $\mathbf{L}(t^\wedge_i)$ (aligned therefore with Q and $\mathbf{L}(t^\wedge_i)$), and for every instant $t^*$, with $t^\wedge_i < t^* < t_2$, $god_{zi}$ remains at a distance $1/z_i$ from $\mathbf{L}(t^*)$ aligned with Q and $\mathbf{L}(t^*)$.

*Proof*: Let $t^\wedge$ be an instant between $t^o$ and $t_2$ at which Q has moved away from $\mathbf{L}(t^\wedge)$ a distance $\delta$, i.e. $|\mathbb{L}(t^\wedge) - \mathbf{L}(t^\wedge)| = \delta$. In symbols:

$$t^o < t^\wedge < t_2 \wedge d_{t^\wedge}(Q, L(t^\wedge)) = \delta \tag{10.13}$$

Since we assumed from the beginning that the World lines of all the particles are always continuous functions of time, we know that $\mathbb{L}(t^o) = \mathbf{L}(t^o)$. From this and (10.13), it follows that:

There is an infinite sequence of positive consecutive integers $z_i$ ($z_1 < z_2 < z_3 < \ldots$), with $1/z_i < \delta$, and an infinite sequence of real numbers $t^\wedge_i$ ($t^\wedge_1 > t^\wedge_2 > t^\wedge_3 > \ldots$) convergent to $t^o$, with $t^o < t^\wedge_i < t^\wedge$, such that $t^\wedge_i$ is an instant between $t^o$ and $t_2$ at which Q has moved a distance $1/z_i$ away from $\mathbf{L}(t^\wedge_i)$, i.e such that $|\mathbb{L}(t^\wedge_i) - \mathbf{L}(t^\wedge_i)| = 1/z_i$.

In symbols:

$$t^o < t_1^{\wedge} < t_2 \wedge d_{t_1^{\wedge}}(Q, \mathbf{L}(t_1^{\wedge})) = 1/z_1$$
$$t^o < t_2^{\wedge} < t_2 \wedge d_{t_2^{\wedge}}(Q, \mathbf{L}(t_2^{\wedge})) = 1/z_2$$
$$t^o < t_3^{\wedge} < t_2 \wedge d_{t_3^{\wedge}}(Q, \mathbf{L}(t_3^{\wedge})) = 1/z_3$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$t^o < t_i^{\wedge} < t_2 \wedge d_{t_i^{\wedge}}(Q, \mathbf{L}(t_i^{\wedge})) = 1/z_i \tag{10.14}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$z_1 < z_2 = z_1 + 1 < z_3 = z_2 + 1 < \ldots$$
$$t_1^{\wedge} > t_2^{\wedge} > t_3^{\wedge} > \ldots \tag{10.15}$$
$$t^o < t_i^{\wedge} < t_2 \tag{10.16}$$
$$lim_{i \to \infty} t_i^{\wedge} = t^o. \tag{10.17}$$

Particularizing (10.11) in the form $(t_1 < \hat{t_i} < t_2 \wedge d_{t\wedge i}(Q, \mathbf{L}(t\wedge_i)) = 1/ z_i) \rightarrow G_{zi}Q_{t\wedge it2}$ and given that, by (10.16), $(t^o < t\wedge_i < t_2 \wedge d_{t\wedge i}(Q, \mathbf{L}(t\wedge_i)) = 1/z_i) \rightarrow G_{zi}Q_{t\wedge it2}$, using modus ponens with (10.14), $G_{zi}Q_{t\wedge it2}$.

We conclude that:

$$\forall i\, G_{zi}Q_{t_i^{\wedge}t2}. \tag{10.18}$$

Expressing this non symbolically: $\forall i$, $god_{zi}$ positions itself at $t\wedge_i$ at a distance $1/z_i$ from $\mathbf{L}(t\wedge_i)$ with Q between it and $\mathbf{L}(t\wedge_i)$ (aligned therefore with Q and $\mathbf{L}(t\wedge_i)$) and, for every instant t*, with $t\wedge_i < t^* < t_2$, $god_{zi}$ remains at a distance $1/z_i$ from $\mathbf{L}(t^*)$ aligned with Q and $\mathbf{L}(t^*)$. ⊣

From (**V**) and (**VI**) it now easily follows that:

(**VII**) If, at the instant $t^o$ between $t_1$ and $t_2$, Q were not to interact in any way with the gods (everything else remaining the same), a massive part of Q would interpenetrate with a massive part of the set of gods at instants of time $t' > t^o$ arbitrarily close to $t^o$.

*Proof*: Our initial assumption that at some instant $t^o$ between $t_1$ and $t_2$ Q does not interact in any way with the gods (everything else remaining the same, i.e. everything compatible with non-interaction between Q and the gods at $t^o$ remaining the same) brings us to a situation where, by (**V**) and (**VI**), Q occupies, for all $t \in (t_1, t_2)$, not the position $\mathbf{L}(t)$ but the position $\mathbb{L}(t)$, characterized (as we saw in (10.14)), by the fact that, at instants $t\wedge_j$, Q is at a distance $1/z_j$ from $\mathbf{L}(t\wedge_j)$, and (as we saw in (10.18)), $\forall i\ G_{zi}Q_{t\wedge it2}$.

Furthermore, since for $t\wedge_i > t\wedge_j$, the distance between Q and $\mathbf{L}(t\wedge_i)$ (namely, $1/z_i$) is greater than the distance between Q and $\mathbf{L}(t\wedge_j)$ (namely, $1/z_j$), we may assume

without loss of generality that, at instants $t^\wedge_i$, Q is at a distance $1/z_i$ from $\mathbf{L}(t^\wedge_i)$ and moving away from $\mathbf{L}(t^\wedge_i)$. In other words, using once again the hypothesis of continuity, for instants of time $t' > t^\wedge_i$ arbitrarily close to $t^\wedge_i$, Q will be at a distance $> 1/z_i$ from $\mathbf{L}(t')$. By $\forall i\ G_{zi}Q_{t^\wedge it2}$, this means that, $\forall i$, a massive part of Q would interpenetrate with a massive part of $god_{zi}$ at instants of time $t' > t^\wedge_i$ arbitrarily close to $t^\wedge_i$. But, since we saw in (10.15), (10.16), and (10.17) that $t^\wedge_1 > t^\wedge_2 > t^\wedge_3 > \ldots$, $t^o < t^\wedge_i$ and $\lim_{i\to\infty} t^\wedge_i = t^o$, it follows that a massive part of Q would interpenetrate with a massive part of at least one god (and, therefore, with a massive part of the set of gods) at instants of time $t' > t^o$ arbitrarily close to $t^o$. ⊣

In view of (**VII**), we may directly apply the Principle of interaction by impenetrability PII. It follows that Q interacts by impenetrability at $t^o$ with the set of gods. Since $t^o$ was an arbitrary instant between $t_1$ and $t_2$, it also follows that, in generalization G2, Q interacts by impenetrability with the set of gods between $t_1$ and $t_2$. This interaction is the one that explains why Q follows the World line $L \equiv \{(t, \mathbf{L}(t)) / t_1 < t < t_2\}$ between the instants $t_1$ and $t_2$. Appealing to the Principle of interaction by impenetrability enables us to understand the origin of the force $\mathbf{F} = m\, d^2\mathbf{L}/dt^2$ the gods exert on Q without having to introduce any "strange field of force."

### *10.3.4 Global Interaction at a Distance in G2*

Let Q, at instant $t^o$ between $t_1$ and $t_2$, not interact in any way with the gods. Although, as we have already seen, it follows from this that a massive part of Q would interpenetrate with a massive part of the set of gods at instants of time $t' > t^o$ arbitrarily close to $t^o$, it by no means follows that a massive part of Q would interpenetrate with a massive part of $god_1$ at instants of time $t' > t^o$ arbitrarily close to $t^o$, or that a massive part of Q would interpenetrate with a massive part of $god_2$ at instants of time $t' > t^o$ arbitrarily close to $t^o$, or that a massive part of Q would interpenetrate with a massive part of $god_3$ at instants of time $t' > t^o$ arbitrarily close to $t^o$, … etc.

This means that, in G2, the effect on Q would be the same in spite of the fact that we would have eliminated one of the gods (indeed, quite clearly, even if we suppressed some finite set of gods). It follows that, in G2, neither $god_1$, nor $god_2$, nor $god_3$, nor…, interacts with Q.

The final conclusion to draw is striking. Although, in generalization G2, Q interacts by impenetrability with the set of gods, it does *not* interact separately with any particular god. Since Q follows the World line $L \equiv \{(t, \mathbf{L}(t))/t_1 < t < t_2\}$ between the instants $t_1$ and $t_2$, $\forall t\ (t_1 < t < t_2 \rightarrow d_t(Q, \mathbf{L}(t)) \neq 1/n)$.

This non-interaction with any particular god, separately from the others, is compatible with what has been stated in the second part of $C_n^{(2)}$, namely that if, at some instant t, with $t_1 < t < t_2$, Q does not happen to be at a distance 1/n from $\mathbf{L}(t)$, then $god_n$ shall not apply any force at all on Q at t:

$$\forall t(t_1 < t < t_2 \rightarrow [d_t(Q, \mathbf{L}(t)) \neq 1/n \rightarrow \neg g_n(Q, t)]).$$

Since no god manipulates Q in any way, all the gods, during the interval of time $(t_1, t_2)$, may remain at a finite distance from Q greater than a given constant value. This implies that the gods interact at a distance with Q (although not individually, as we already know, but globally).

This is not a paradox; it follows from the possibility that the gods move as fast as necessary in order to interfere with the evolution of Q. What G2 shows is that, if we allow interaction by contact and we place no limit whatsoever on the velocity that a material object can attain, interaction at a distance becomes a genuine possibility. It corresponds to the "strange field of force" mentioned by Benardete.

Some years ago, in Laraudogoitia (2005), I examined another complementary example of a "strange field of force": if we permit interaction by contact and we place no limit whatsoever on the length that a material object may have, then interaction at a distance becomes a possibility. I considered the case of a flat, infinite surface on which lay two identical infinite javelins (perfectly rigid), one being placed parallel to the other at a unit distance. If Achilles makes the javelin to his left rotate clockwise (without lifting it off the ground), the one to his right will also simultaneously begin to rotate clockwise, although the distance separating the two is finite at all times. There is no contradiction in this case either with the laws of classical mechanics: "It is important to observe that classical mechanics allows us to determine in detail the future evolution of the two javelins depending on the external forces that Achilles exerts on one of them" (Laraudogoitia op. cit.: 437).

One important difference still remains: while a purely mechanical model is obtained from the paradox of Achilles' javelin, I have merely given a model that is compatible with classical mechanics from the generalization of the Benardete paradox (see, however, Sect. 10.6 below). The central point of the analogy is nevertheless clear: the mere possibility of interaction by contact may in certain conditions give rise to interaction at a distance. Interaction at a distance between the javelins is a direct consequence of the possibility of their interaction by contact (which requires that the javelins, as material bodies, should not mutually interpenetrate) and of the procedure Achilles follows in acting on one of them (without interaction at a distance the javelins would interpenetrate mutually as Achilles only makes one of them rotate). Analogously, (arbitrarily complex) interaction at a distance between the set of gods and Q in G2 is a direct consequence of the possibility of their interaction by contact and of the action procedure studied in my generalization of the Benardete paradox (which presupposes this possibility). One of the purposes of the paper was indeed to make this point clear.

## 10.4 The Physics Underlying the Paradox (II)

Neither Benardete's formulation nor my generalization of it lead to a contradiction since all they state is that the gods stop the particle or, in my generalization, that they move it along the World line L. They do not attempt to explain how they manage to influence its movement, which is irrelevant with regard to the logical structure of the paradox.

Going one step further, we saw in the previous section that the gods interact globally by impenetrability with particle Q, but this is not sufficient to guarantee the physical coherence of the situation. A problem appears when one considers the means by which the gods affect the particle. In G2, the gods apply a total force $m\ d^2\mathbf{r}/dt^2 = m\ d^2\mathbf{L}/dt^2$ on Q, but from $C_n^{(2)}$ it follows that no god exerts any force on Q. How is it possible that, if no god applies any force on Q, the set of gods can do so?

This, in my view, is the main problem underlying the Benardete Paradox of the gods. An equivalent form of this difficulty can be seen by resorting to the principle of action and reaction (ARP), which may be formulated thus: the force that a physical system X exerts on a physical system Y is equal and opposed to the force that Y exerts on X, and both forces are contained in the same straight line (see, for example, Goldstein 1959).[6] In G2, the gods exert a force $m\ d^2\mathbf{r}/dt^2 = m\ d^2\mathbf{L}/dt^2$ on Q, such that Q must exert a force $-m\ d^2\mathbf{L}/dt^2$ on the set of gods. But none of the gods exerts any force on Q, so Q exerts no force on any of them. How is it possible for Q to exert a non-null force on the set of gods without exerting any force on any of the gods separately? This is not a question one may address by merely appealing to interaction by impenetrability. Although that does not mean that any specific model of G2 will be physically impossible, it is not hard to find models in which that physical impossibility occurs.

### *10.4.1 G2(I): A Physically Impossible Model of G2*

Let us consider the following model of G2, which I shall call G2(I). Let us suppose that, $\forall t \leq t_1 = 0$, particle Q is at rest at x = 0, taking $t_2 = +\infty$. Let the World line L for $0 < t < \infty$ correspond to a uniformly accelerated movement of constant positive acceleration U along the axis OX. Let us also suppose that the gods are situated on the axis OY (perpendicular to OX) at the points of Cartesian coordinate (x,y) = (0, n). Since the force on Q is mU and is directed in the positive direction of the axis OX, an equal and contrary force (i.e. in the negative direction of the OX axis) must act, by ARP, on the set of the gods. But this is impossible, because we supposed that none of the $god_n$ is situated on the OX axis and ARP demands that the force of action and the force of reaction be contained in the same straight line. G2(I) is physically impossible to the extent that ARP is violated (in this paper I assume the underlying physical theory to be formed by Newton's three laws of dynamics), but

there is no logical impossibility (indeed, as we know, ARP is also violated, e.g., in classical electrodynamics).

In a world governed by laws that include ARP (as is the case of a world governed by Newton's three laws of dynamics), gods attempting to follow the plan specified by the $C_n^{(2)}$ in G2(I) will fail. This means that not all $C_n^{(2)}$ may be true (further, an infinite number of $C_n^{(2)}$ will be false, as is immediately clear). But the gods will fail for physical, not for logical, reasons. The Benardete paradox and its generalizations may, in some of their versions, imply the presence of forms of interaction between the gods and the particle Q that happen to be incompatible with specific physical laws (e.g. with ARP).

## 10.5 Circumventing the Problem with ARP

I claimed that G2 is not in principle incompatible with ARP (unlike G2(I), which clearly is), but it remains unclear how one could construct a model in which the compatibility would be made explicit (i.e. a model in the Benardete spirit, involving a numerable infinity of gods). I now address this particular task, which will also have other interesting consequences. The idea is to construct a generalization of the Benardete paradox of the gods (another model of G2) that is explicitly compatible with ARP.

### *10.5.1 G2(II): A Physically Interesting Model of G2*

Let us suppose from the outset that each $god_n$ has control over a wall $w_n$ with which it is possible to manipulate a particle by contact, i.e. "by pushing." Let us stipulate that, at the instant $t = t_1 = 0$, we have particle Q (of mass m) at rest on the axis OX at point $x = +h$, while the walls are all at rest and in mutual contact, occupying the region between $x = 0$ and $x = -k$ of the x axis, so that wall $w_n$ (which may be manipulated by $god_n$) occupies the part of this region between $x = -k/n$ and $x = -k/(n + 1)$. We shall call $L_{0,+h} \equiv (t, \mathbf{L}_{0,+h}(t))$ the World line of any particle that, starting from rest at $x = +h$ in $t = 0$, moves with increasing acceleration a(t) towards the right (i.e. the positive region of the X axis). We shall now consider the following model of G2, which I call G2(II):

$C_n^{(2(II))}$: if at some instant t, with $0 < t < +\infty$, Q comes to be at a distance 1/n from $\mathbf{L}_{0,+h}(t)$, then for every instant t*, with $t < t^* < +\infty$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from $\mathbf{L}_{0,+h}(t^*)$, making wall $w_n$, which it controls, interact by contact with Q for that purpose. If, at some instant t, with $0 < t < +\infty$, Q does not come to be at a distance 1/n from $\mathbf{L}_{0,+h}(t)$, then $god_n$ shall not apply any force at all on Q at t.

If we introduce:

C: particle Q follows the World line $L_{0,+h}$ between the instant $t_1 = 0$ and $t_2 = +\infty$.

then, given that $(\wedge_n C_n^{(2(II))})$ is fulfilled, C follows, i.e. particle Q follows the World line $(t, \mathbf{L}_{0,+h}(t))$ for $t \geq 0$. Since Q does not diverge from the World line $L_{0,+h}$ between 0 and $+\infty$, no god exerts any force on Q in that interval. Even so, in case Q did not interact with the walls and, therefore, diverged from the World line $L_{0,+h}$ between 0 and $+\infty$, Q would interpenetrate with the walls, which indicates that the force $F(t) = m\,a(t)$ that acts on Q for $t > 0$ is applied by the walls, this being an interaction by impenetrability. It is the walls indeed that will now, if necessary, enter into direct contact with Q, just as the gods themselves did in the case considered in the discussion of Sect. 10.3. Thus, the same argument that led us to global interaction by impenetrability *of the gods* with Q leads now to global interaction by impenetrability *of the walls* with Q.

Furthermore, the interaction at a distance by impenetrability of Q with the walls is compatible with ARP. As a matter of fact, the initial spatial arrangement of the walls was designed to allow for a simple description of the result of this global interaction: by being all together in mutual contact in the way specified previously, the walls will move "*en bloc*" to the left ($w_1$ being pushed by $w_2$, $w_2$ by $w_3$, and so on) with the same linear momentum in absolute value as Q's, but towards the opposite sign, being thus compatible with Newton's laws and, in particular, with ARP. No physical law is violated (i.e. none of Newton's three laws), and the gods have triggered the increasing separation between Q and the set of walls through their interaction by impenetrability (i.e. through the interaction by impenetrability between Q and the set of walls).

## 10.6 A Purely Mechanical Model of the Benardete Dichotomy

In what follows, our hitherto thoroughly general discussion will be complemented by an analysis of the paradox of the gods construed as a particular case; among other things, this will enable us to check which role the Principle of interaction by impenetrability (PII) plays in a specific and particularly simple case.

Let us suppose that each $god_n$ has control over a wall $w_n$ with which it is possible to manipulate (by contact) an object (a man). The following condition is quite faithful to the gist of the paradox of the gods in Benardete's original formulation:

(a) $god_n$ prevents the man from going further than $1/2^n$ miles if and only if the man actually goes $1/2^n$ miles

which, making the walls intervene explicitly, we may rewrite thus:

(b) $wall_n$ prevents the man from going further than $1/2^n$ miles if and only if the man actually goes $1/2^n$ miles.

### 10.6.1 The Model

In order to construct a purely mechanical model of this situation, we shall replace the man with a particle Q moving towards the right at unit velocity on the negative region of the X axis. The role of $w_n$ will be taken by a particle $P_n$, centered and at rest at point $x = 1/2^n$. Let us suppose that all the particles involved have the same unit mass; although they are not point particles, for the sake of clarity we shall not take their size into consideration in the discussion. An elementary result of classical dynamics (as well as of relativistic dynamics) establishes that the collision between a particle A, moving at velocity v, and another particle B at rest of identical mass, results in A being at rest and in B moving at velocity v (provided we are dealing with a movement in a single spatial dimension), i.e. with B preventing the subsequent progress of A.

Applying this idea to the case of the pair of particles Q and $P_n$:

(c) if Q reaches the point $x = 1/2^n$, $P_n$ prevents it from going beyond the point $x = 1/2^n$.

But, understanding the phrase "prevents it from going beyond" in terms of a physical action, it is an analytical truth that:

(d) if Q does not reach the point $x = 1/2^n$, then it is not true that $P_n$ prevents it from going beyond the point $x = 1/2^n$.

From (c) and (d) we then conclude that:

(e) $P_n$ prevents Q from going beyond point $x = 1/2^n$ if and only if Q reaches point $x = 1/2^n$.

The strict formal analogy between (a), (b) and (e) guarantees that we have indeed constructed a purely mechanical model of the paradox of the gods. On this basis, checking which role (PII) plays turns out to be an easy task.

### 10.6.2 The Role of Impenetrability

Let us suppose that Q reaches point $x = 0$ at $t = t^o$. If, at this instant, Q did not interact in any way with the particles $P_n$ so that the force on it were 0 (everything else remaining the same, i.e. everything being compatible with non-interaction between Q and the particles $P_n$ at $t^o$), the velocity of Q at $t = t^o$ would still be the unit and, therefore, for instants of time $t' > t^o$ arbitrarily close to $t = t^o$, Q would be in the region $x > 0$. But this means that Q would have interpenetrated with at least one particle $P_n$ (and therefore with a massive part of the set of particles $P_n$) at instants of time $t' > t^o$ arbitrarily close to $t = t^o$.[7] We conclude that if, at instant $t^o$, Q were not to interact in any way with the particles $P_n$ (everything else remaining the same), a

massive part of Q would interpenetrate with a massive part of the set of particles $P_n$ at instants of time $t' > t^o$ arbitrarily close to $t = t^o$.

Applying (PII), it follows that Q interacts by impenetrability at $t = t^o$ with the set of particles $P_n$. Thus, PII enables us to understand the origin of the force that the particles $P_n$ exert on Q without having to introduce any "strange field of force."

Let Q at instant $t^o$ not interact in any way with the particles $P_n$. Although, as we have already seen, it follows from this that a massive part of Q would interpenetrate with a massive part of the set of particles $P_n$ at instants $t' > t^o$ arbitrarily close to $t^o$, it by no means follows that a massive part of Q would interpenetrate with a massive part of $P_1$ at instants $t' > t^o$ arbitrarily close to $t^o$, or that a massive part of Q would interpenetrate with a massive part of $P_2$ at instants $t' > t^o$ arbitrarily close to $t^o$, or that a massive part of Q would interpenetrate with a massive part of $P_3$ at instants $t' > t^o$ arbitrarily close to $t^o$, …, etc. This means that, in my mechanical model of the paradox of the gods, the effect on Q would be the same despite the fact that we would have deleted one of the particles $P_n$; it is clear indeed that it would be the same even if we had deleted any finite set of particles $P_n$. It follows that neither $P_1$, nor $P_2$, nor $P_3$, nor … interacts in any way with Q. Although Q interacts by impenetrability with the set of particles $P_n$, it does not interact separately with any particular particle $P_n$.

We must now check that my mechanical model of the paradox of the gods is also compatible with the laws of conservation of mechanics (the relevant ones here, as in the usual processes of collision in one-dimensional motions, are the law of conservation of momentum and the law of conservation of energy). This compatibility follows automatically if we suppose, for instance, that at every instant one and only one of the particles involved in the model has a unit velocity (except, of course, at the instants at which a collision takes place). Although this is not the only possible supposition warranting compatibility (so that the evolution of the system is in fact indeterminist, something by no means difficult to demonstrate), it is indeed the simplest.[8] The interaction (collision) of Q with the set of $P_n$ takes place at $t^o$ and, before this instant, Q is the particle with unit velocity. How is it possible that there always be one, and only one particle with unit velocity? Let $t_j$ be an instant after $t^o$ in which $P_j$ has velocity $v_j = 1$. $P_j$ was set in motion by the collision with $P_{j+1}$, which was in movement ($v_{j+1} = 1$) at certain instants $t_{j+1}$ ($t^o < t_{j+1} < t_j$) and was set in motion by the collision with $P_{j+2}$, which was in movement ($v_{j+2} = 1$) at certain instants $t_{j+2}$ ($t^o < t_{j+2} < t_{j+1}$) … and so on. So it is clear that $P_j$ ($j > 1$), which was originally at rest at $x = 1/2^j$, acquired velocity $v_j = 1$ from its collision with $P_{j+1}$ and transferred it (returning to rest) to $P_{j-1}$. In other words, the system of $P_j$ ($j \geq 1$) has become excited at $t^o$ (i.e. $\lim_{j\to\infty} t_j = t^o$) as a consequence of its global interaction (collision) with Q at that instant, both momentum and kinetic energy being conserved throughout the process.

In my mechanical model, the global interaction between Q and the set of $P_n$ took place at an instant $t^o$ in which Q's distance from the set was null (namely, when Q was at point $x = 0$). But it is not essential that this distance be null at the instant at which the global interaction takes place. To see why, let us suppose that, initially (for $t < t^o$, just as before), Q is moving towards the right on the negative region of

the X axis at unit velocity while the particles $P_n$ are centered and at rest, not at points $x = 1/2^n$, but rather, say, at points $x = 1+ (1/2^n)$. If we suppose no more than this, we already know that Q would collide globally with the set of $P_n$ at $x = 1$ at instant $t^o+ 1$. But suppose that we make Benardete's gods intervene in this scenario in the following way. $God_n$ controls particle $P_n$ and decides to prevent Q from moving beyond point $x = 1/2^n$ by placing the particle $P_n$ there (after removing it from its initial location at $x = 1+ (1/2^n)$) in case Q reaches that point. Let us say that $god_n$ removes $P_n$ from its initial location at $x = 1+ (1/2^n)$ and quickly places it at $x = 1/2^n$ if $god_n$ sees Q reach the mid point between $x = 1/2^{n+1}$ and $x = 1/2^n$.

Obviously it is still true that:

(e) $P_n$ prevents Q from going beyond the point $x = 1/2^n$ if and only if Q actually reaches the point $x = 1/2^n$.

The previously deduced conclusion, to the effect that Q interacts globally with the set of $P_n$ at instant $t^o$ at point $x = 0$, so that it cannot pass beyond that point, may therefore be maintained. By being unable to pass beyond it, it is clear that Q will not reach any point $x = 1/2^n$, which means that no $god_n$ will remove the particle $P_n$ it controls from its initial location at $x = 1+ (1/2^n)$. This means that Q interacts globally at a distance with the set of $P_n$ at point $x = 0$, being at a unit distance from the set and not in contact with it. This follows from (PII), which we explored in a more general setting in Sect. 10.3. (Note that, unlike what happens in the present case, the conservation of kinetic energy could not then be guaranteed, which clearly indicates that my generalization is not a trite variation on the present case). In any event, it is clear that what causes particle Q to be unable to go beyond point $x = 0$ is its interaction with the set of particles $P_n$. The intentions of the gods are not the cause.

### *10.6.3 Reply to Yablo-Type Criticisms*

Taking advantage of the simplicity of my mechanical model, I shall now resort to it to criticize one of the best known diagnoses of the paradox of the gods, a diagnosis that is also one of the most influential for anyone who has cared to explore the subject. I refer to the analysis offered by Yablo in Yablo (2000). The thesis, diametrically opposed to mine, is that the paradox of the gods should be analyzed as a logical problem and not as a physical one. Yablo's idea is that, if Benardete's man moves from A to B, the gods are faced with a task that is logically impossible to perform. Yablo claims that:

> If there's a paradox here, it lies in the difficulty of combining individually operational subsystems into an operational system. But is this any more puzzling than the fact that although I can pick a number larger than whatever number you pick, and viceversa, we can't be combined into a system producing two numbers each larger than the other?
>
> Yablo (2000: 151)

For a detailed understanding of Yablo's argument and of the reason why it is fallacious, we need to return to (e), which implicitly includes a universal quantification on n. A logical consequence of (e) is that:

(f) $\forall n$ (If Q reaches the point $x = 1/2^n$ and $\neg \exists m > n$ such that $P_m$ prevents Q from going beyond point $x = 1/2^m$, then $P_n$ prevents Q from going beyond point $x = 1/2^n$).

But, given the conditions of the situation in my mechanical model, it is clear that:

(g) $\forall n$ (If $P_n$ prevents Q from going beyond point $x = 1/2^n$, then Q reaches point $x = 1/2^n$ and $\neg \exists m > n$ such that $P_m$ prevents Q from going beyond point $x = 1/2^m$)

although (g) is *not* a logical consequence of (e).

Finally, from (f) and (g) we infer that:

(h) $\forall n$ ($P_n$ prevents Q from going beyond point $x = 1/2^n$ if and only if Q reaches point $x = 1/2^n$ and $\neg \exists m > n$, such that $P_m$ prevents Q from going beyond point $x = 1/2^m$).

The conclusion to be drawn is that (h) depends logically on both (e) and (g).

Now, let Q go beyond point x = 0. We may suppose without loss of generality that Q reaches successively all the points $x = 1/2^n$ and goes beyond them. From this and (h) it follows that:

(i) $\forall n$ ($P_n$ prevents Q from going beyond point $x = 1/2^n$ if and only if $\neg \exists m > n$ such that $P_m$ prevents Q from going beyond point $x = 1/2^m$).

But it is logically impossible to carry out (i) since it is a logical contradiction. It would then seem that if Q goes beyond point x = 0, the particles $P_n$ must carry out a logically impossible task.

In the terminology of Benardete's original paradox, one would say that if the man goes from A to B we would analogously infer that:

(j) $\forall n$ ($god_n$ prevents the man from going beyond $1/2^n$ miles if and only if $\neg \exists m > n$ such that $god_m$ prevents the man from going beyond $1/2^m$ miles).

Yablo concludes that, in such a case, $god_n$ would have to carry out a task that it is logically impossible to bring to completion (Yablo op. cit.). This conclusion is however mistaken.

Let us briefly take a look at what happens within my mechanical model. If Q goes beyond point x = 0, reaching successively all points $x = 1/2^n$ and going beyond them, then (e) is clearly false (no $P_n$ prevents Q from going beyond point $x = 1/2^n$). Now if (e) is false, since (h) is deduced from both (e) and (g), there is no justification left for (h) and, therefore, no justification left for (i). The conclusion to the effect that if Q goes beyond point x = 0, the particles $P_n$ must carry out a logically impossible task is therefore unjustified. Indeed (as we have already seen in

Sect. 10.6.2), if Q goes beyond point x = 0, Q will interpenetrate with the set of particles $P_n$, so that they will not comply with (c) nor, therefore, with (e). In the terminology of Benardete's original paradox, the argument to the effect that, if the man goes from A to B, $god_n$ must carry out a logically impossible task is fallacious. In other words, there is no reason to think that the man must remain at point A in order for the infinite set of all of the gods' intentions to be consistent. What one should say is that the man remaining at point A is a necessary condition for the fulfilment of an infinite number of the gods' intentions. This underwrites my interpretation of the paradox of the gods and its generalizations as a physical problem rather than as a logical problem. The only thing one may reasonably maintain from a purely logical perspective is that, if the man went from A to B, the gods would not fulfill their intentions. Sentences (a) and (b) would simply be false just as, in my mechanical model, (e) would be false in case Q goes beyond point x = 0, reaching successively all the points $x = 1/2^n$ and going beyond them.

But to claim that, if the man went from A to B, the gods would not act according to their intentions, is about as trivial as to say that if my intention were to start my car's engine today at midday, and it turned out that my car's engine didn't start today at midday, I would not have fulfilled my intention.

## 10.7 Introducing Subtasks

In all the generalizations of the Benardete paradox of the gods that we have looked at in the previous sections (and something similar could be said of Benardete's own formulation), the clauses $C_n^{(1)}$, $C_n^{(2)}$, and $C_n^{(2(II))}$ describe vacuously performed conditioned actions in the precise sense that none of their antecedent conditions are fulfilled (although $C_n^{(2)}$ and $C_n^{(2(II))}$ also describe conditioned omissions of actions whose antecedent conditions are fulfilled). In this respect, the generalization G1 was particularly interesting. The proof of $C \leftrightarrow (\wedge_n C_n^{(1)})$ made it clear that any sentence describing a possible physical process (typically exemplified by the possible process described by sentence C) is equivalent to (the conjunction of) an infinite set of sentences describing vacuously performed conditioned actions. In other words, any possible physical process may be equivalently described as an infinite set of vacuously performed conditioned actions or, to put it yet more tersely, any possible physical process is indeed equivalent to an infinite set of vacuously performed conditioned actions.

However, although the evolution of particle Q between $t_1$ and $t_2$ is a possible physical process, it may occur in the context of a broader physical system whose evolution between $t_1$ and $t_2$ is not itself a possible physical process. This is what some of the generalizations of Benardete's paradox we've been considering show. For example, in G2(I) the gods act on particle Q by applying the force mU on it without, in return, any equal and contrary force acting on the gods. In this case, the evolution of the broader physical system formed by particle Q and the gods is obviously not an example of a possible physical process because, as we saw above,

ARP is violated. Furthermore, generalization G2(II) shows that the evolution of the broader physical system formed by particle Q and the walls is not an example of a conventional process[9] despite the fact that it is a physically possible process (in the sense, assumed in this paper, of compatibility with Newton's three laws of movement). This shows that in many cases such as G2(II), infinite sets of vacuously performed conditioned actions imply processes that, while not being conventional, are nevertheless physically possible.

In view of their theoretical relevance, I suggest we call such sets "subtasks," by analogy with the term "supertask" applied to infinite sequences (sets) of actions. In other words, a subtask is an infinite set of vacuously performed conditioned actions.[10] The reference to "actions" in this context is in no way restrictive because one may always consider nature itself to be an actor, in which case subtasks and supertasks will deal simply with physical processes in general.[11]

There are relevant similarities between the two concepts that need to be explored in more detail in the future. For instance, we have seen that any physical process may be equivalently described as a subtask, and we likewise know that any physical process may be equivalently (and trivially) described as a supertask.[12] When Thomson introduced the concept of supertask (see Thomson 1954–1955), the idea had already matured in the literature published around that time (see, e.g., Black 1950–1951; Grünbaum 1955; Weyl 1949) and the new concept helped to unify and systematize an entire branch of thinking about the infinite. I believe that a new sphere of reflection on the infinite, which began with Benardete 1964 and has been considered until now in several publications (see, e.g., Priest 1999; Shackel 2005; Hawthorne 2000; Peijnenburg and Atkinson 2010 and Uzquiano 2012), may also be unified and systematized under the concept of subtask I have proposed.[13]

## Appendix: The Way to Subtasks

A supertask is an infinite sequence of actions (processes) carried out (taking place) in a finite time. A task is a finite sequence of actions (processes). It must necessarily be carried out (take place) in a finite time because, by definition, the carrying out of an action only lasts a finite time. Hereafter, I use the terms "actions" and "processes" indifferently as befits the case. The notions of supertask and task may be formulated in terms of sets of actions instead of in terms of sequences of actions without significant change. Along these lines, we may try to define a subtask as:

(a) an infinite set of actions not carried out in a finite time.

The idea is that there exists a certain finite interval of time in which none of the actions of the set has been carried out. But the fact that actions haven't been brought to completion in a finite interval of time is no more controversial than their non-completion in an infinite interval of time. Considered from another perspective, time is relevant when discussing actions that have indeed been carried out (and

these lead to changes in time), but it doesn't appear to be so when no action is performed at all. This suggests that we should leave time out of the description.

If we defined a subtask as:

(b) an infinite set of actions not carried out

subtasks would be uninteresting in a vast majority of cases, except perhaps those in which there are conditions for actions to be carried out. If these conditions are sufficient conditions then we know, that, by definition, they will not be carried out in a subtask. Thus, a subtask would be an infinite set of not carried out actions, with non-carried out sufficient conditions for their completion. In other words, a subtask would be defined as:

(c) an infinite set of actions whose carrying out (which happens not to take place) is (materially) conditioned by the realization of certain states of affairs which do not take place either.

Notice that if we do not impose the condition that the states of affairs should be relevant, it follows that, since (as a consequence of the properties of the material conditional) any falsity would be a sufficient condition (as well as a necessary one) for the carrying out of some action in a subtask, (c) would be extensionally equivalent to the definition of a subtask construed as a an infinite set of actions (of processes) whose carrying out does not take place.

But in that case, we would end up with an unnecessarily broad notion (as in the case of (b)). In Sect. 10.7, I defined a subtask as an infinite set of vacuously performed conditioned actions (processes). In more explicit terms, and as a result of the necessary modification to (c), the definition now takes this form:

(d) a subtask is an infinite set of actions whose carrying out, which does not take place, is (materially) conditioned to the realization of certain relevant states of affairs that do not take place either.

This suggests that it would be suitable (by analogy) to extend the definition of supertask so that an infinite *set* of actions, whose completion takes place during a certain interval of finite time, is now taken into account in the definition. In this case, an infinite set of actions whose completion is (materially) conditioned by the realization of certain relevant states of affairs could, in principle, either be a subtask (in case the actions end up not being carried out and the states of affairs end up not being realized), or be a supertask (in case infinite actions are not only brought to completion, but brought to completion in a finite time). A subtask is carried out although none of the infinite actions required by its definition is brought to completion, and none of its relevant states of affairs is realized. By contrast, a supertask is carried out whenever infinite actions are brought to completion (in a finite time), and a task is carried out whenever only a finite number of actions are brought to completion. With respect to the carrying out of actions, a supertask goes beyond a mere task, but a subtask does not even achieve the status of a task, hence its name.

As I suggested above, a subtask will only be non-trivial in case, by way of relevant states of affairs, there are sufficient, or necessary and sufficient conditions, of a certain type. Any infinite set of actions (processes) whose carrying out does not take place would in principle allow us to define a subtask because, in such a case, any infinite set of actions may be (materially) conditioned (at the very least if we take into account sufficient conditions) by the realization of relevant states of affairs that are not realized (and which, furthermore, can be realized in infinitely many ways). The notion of subtask (just like the notion of supertask) therefore includes a multitude of realizations of no interest, as I already noted above. But there are nevertheless interesting cases. An interesting supertask is one that has *non*-trivial consequences, and the very same thing may be said of interesting subtasks. Although the notion of "trivial" is unmistakably vague, the vagueness merely reflects that of the notion of what is deemed "interesting."

In any event, (d) provides the concept of subtask with some structure, by allowing it to have non-trivial consequences; this is indeed a progress over the non-structured (b). We're now facing something which is indeed more interesting than supertasks because it seems a priori impossible for the carrying out of a subtask (as defined as in (d)) to have significant consequences. Just as one must provide an infinite sequence of carried out relevant actions when describing a supertask, one, when describing a subtask, must provide an infinite set of relevant actions (of relevant processes), whose carrying out does not take place, and a set of relevant conditions (relevant states of affairs) for its carrying out, which are not realized either. In many cases, it will be sufficient, given the situation, to specify the relevant conditions in order to deduce that none of the actions of the infinite set of relevant actions will be carried out. This is exactly what we have seen in the paradox of the gods and its generalizations.

I would like to conclude with a note on the terminology. I have talked of "carrying out" a subtask despite the fact that one necessary condition for being a subtask is that none of the infinite actions that play a role in its definition is actually carried out. My intention was to underline the formal parallels with the concept of supertask. It would perhaps be more suitable to talk of carrying out* a subtask, analogously to how some authors use the term "causation*" to describe cases of prevention or omission in causal terms. The expressions "observe" or "comply with" (as when one says that one "observes the rules" or "complies with the law") might appear to be a suitable alternative to "carry out" in the context of subtasks. In the absence of a clear alternative, I have not followed any of these paths. My purpose was to point out the relevance of the concept of subtask.

## Notes

1. $C_n^{(1)}$ is a universal sentence, namely, $C_n^{(1)}$: for every instant t, with $t_1 < t < t_2$, if Q, at t, comes to be at a distance 1/n from $\mathbf{L}(t)$, then for every instant t*, with $t < t^* < t_2$, $god_n$ will ensure that, at t*, Q is at a distance $\leq$ 1/n from $\mathbf{L}(t^*)$.
2. Since all the material bodies I consider in this paper are assumed to be non-point, Q and $god_n$ (in particular) are non-point, which means that they cannot be at the same distance from $\mathbf{L}(t)$ and, at the same time, that Q be

between **L**(t) and $god_n$. The small changes required to put this right are trivial and I do not take them into consideration here because they would complicate the argument unnecessarily.

3. In other words, everything that is compatible with non-interaction between them at t remains the same.
4. There are a number of philosophical subtleties concerning contact and impenetrability that I do not take into account in this paper. My stance on this point is in line with the position defended by Smith, for whom such subtleties have little importance within the framework of the mechanics of continuous media that one finds as an essential chapter of classical physics (see Smith 2007). The principle of interaction by impenetrability (PII) isn't one of such subtleties. What it does, as the example of particle Q and cylinder C shows, is to facilitate an explanation of why two rigid bodies interact by collision when they do (and do so, furthermore, according to the relevant laws of conservation): PII takes physics seriously.
5. I do not explore here questions of differentiability, which would require a more sophisticated treatment, but wouldn't, in any event, compromise the validity of (**V**).
6. The principle of action and reaction assumes that forces are always given between two bodies (physical systems) such that the total force on a body is always the sum of two-to-two interactions. As Margenau makes clear, if the force between two molecules depends on the distance between them, and molecule 3 introduces a difference in the force of 2 on 1, the relative position of molecules 2 and 1 remaining unchanged, we are faced with a force between three bodies (see Margenau 1950).
7. Indeed. For $t \leq t^o$, the center of mass of the total system of infinite particles Q, $P_1$, $P_2$, $P_3$, $P_4$, … is the point x = 0 (remember that they all have the same mass). Now let Q be at $t = t^o+\varepsilon$ at point $x = \delta > 0$. If it has not interpenetrated with any particle $P_n$, this can only be due to all of these particles remaining to its right. But since in $t^o$ there were infinite particles $P_n$ between x = 0 and $x = \delta$, they must have moved in the interval of time $\Delta t = \varepsilon$ to the right of the point $x = \delta$. Therefore, the center of masses of the complete system must now be at a point $x \geq \delta$. This displacement of the center of masses of the isolated system formed by the total set of particles Q, $P_1$, $P_2$, $P_3$, $P_4$, ….. goes against Newton's 1st law of movement (the law of inertia), which implies that if Q is at $t = t^o+ \varepsilon$ at point $x = \delta > 0$ (in a more general form, if for instants of time $t' > t^o$ arbitrarily close to $t = t^o$ Q is in the region x > 0), then it must have interpenetrated with at least one particle $P_n$ (and, therefore, with a massive part of the set of particles $P_n$) at instants $t' > t^o$ arbitrarily close to $t = t^o$.
8. Alternatively (and rather more elegantly), one might suppose that at no moment shall any of the particles involved move to the left (in the negative direction of the X-axis). Then it is easy to prove that the requirement of the conservation of kinetic energy and momentum leads, at every instant, to one and only one of the particles involved in the model having unit velocity (except, of course, at the instants at which a collision takes place).

9. By "conventional process," I mean a process with interaction by contact and in accordance with Newton's laws of movement.
10. To appreciate the relation between subtasks and supertasks more closely, it is interesting to note that an infinite set of non-vacuously carried out conditioned actions is simply an infinite set of actions *simpliciter* and, therefore, a supertask. In order to have a subtask, the conditioned actions must be vacuously carried out. Note also that in subtask G2(II), the principle of conservation of energy would appear to be violated (given that the kinetic energy of the system formed by the walls and Q grows constantly), something that is likewise characteristic of many standard examples of dynamic supertasks (see, e.g., Atkinson 2007).
11. The idea of nature as actor is not that remote from physical theory proper and has been frequently associated with a teleological formulation of the laws. The paradigmatic case in classical mechanics is Hamilton's principle of minimum action. All of this is admissible provided it is clear that, as Mittelstaedt and Weingartner note: "Physical nature has no goals in the sense that living organisms have goals" (Mittelstaedt and Weingartner 2005: 144). In the case of subtask G2(II), if we substitute "the gods" by "Nature," we have a suggestive means of explaining the interaction at a distance between the walls and Q based on the "mechanism" of impenetrability.
12. At least in Newtonian worlds which, as I said above, are the only ones I have taken into consideration in this paper.
13. The parallels with the history of supertasks are also striking. Authors such as, e.g., Thomson (1954–1955) and Black (1950–1951) have argued against the possibility of supertasks; authors such as Yablo (2000), Shackel (2005) and Peijnenburg and Atkinson (2010) have also refused to accept the possibility of subtasks (using different arguments). Although the critical analysis of such arguments was not the objective of this paper (with the exception of Yablo 2000), I think my generalizations of the Benardete paradox may be taken as a useful starting point for a reply to their objections.

**Acknowledgements** Research for this work is part of the research project FFI2015-69792-R (MINECO/FEDER)

## References

Atkinson, D. (2007). Losing energy in classical, relativistic and quantum mechanics. *Studies in History and Philosophy of Modern Physics, 38*, 170–180.

Benacerraf, P. (1962). Tasks, Supertasks and the Modern Eleatics. *The Journal of Philosophy*, LIX, 765–784.

Benardete, J. A. (1964). *Infinity: An essay in metaphysics*. Oxford: Clarendon Press.

Black, M. (1950–1951). Achilles and the tortoise. *Analysis*, *11*(5), 91–101.

Goldstein, H. (1959). *Classical mechanics*. Reading, Mass: Addison-Wesley Publishing Company.

Grünbaum, A. (1955). Modern science and the refutation of the Paradoxes of Zeno. *The Scientific Monthly*, *LXXXI*, 234–239.

Hawthorne, J. (2000). Before effect and zeno causality. *Noûs, 34*(4), 622–633.
Laraudogoitia, J. P. (2005). Achilles' Javelin. *Erkenntnis, 62*(3), 427–438.
Margenau, H. (1950). *The nature of physical reality: A philosophy of modern physics*. New York: McGraw-Hill.
Mittelstaedt, P., & Weingartner, P. A. (2005). *Laws of nature*. Berlin: Springer.
Peijnenburg, J., & Atkinson, D. (2010). Lamps, cubes, balls and walls: Zeno problems and solutions. *Philosophical Studies, 150*(1), 49–59.
Priest, G. (1999). On a version of one of Zeno's paradoxes. *Analysis, 59*(1), 1–2.
Shackel, N. (2005). The form of the Benardete dichotomy. *The British Journal for the Philosophy of Science, 56*(2), 397–417.
Smith, S. R. (2007). Continuous bodies, impenetrability, and contact interactions: The view from the applied mathematics of continuum mechanics. *The British Journal for the Philosophy of Science, 58*(3), 503–538.
Thomson, J. F. (1954–1955). Tasks and super-tasks. *Analysis*, *15*(1), 1–13.
Uzquiano, G. (2012). Before-effect without zeno causality. *Noûs, 46*(2), 259–264.
Weyl, H. (1949). *Philosophy of mathematics and natural science*. Princeton, New Jersey: Princeton UP.
Yablo, S. (2000). A reply to new Zeno. *Analysis, 60*(2), 148–151.

# Chapter 11
# Supertasks, Physics and the Axiom of Infinity

**Antonio León-Sánchez and Ana C. León-Mejía**

## 11.1 Introduction

It seems reasonable to assume that mathematical infinity was not the objective of Zeno's dichotomy (in any of its variants); however, some kind of mathematical infinity was already at stake in his celebrated arguments. Aristotle proposed a solution to Zeno's dichotomy by introducing what we now call one-to-one correspondences, the key instrument of modern infinitist mathematics. But Aristotle, more a naturalist than a platonist, finally rejected the method of pairing the elements of two infinite collections (in the case at hand, points and instants) and introduced instead the distinction between actual and potential infinities. Aristotle's distinction served to define two opposite positions on the nature of infinity for more than twenty centuries. Actual infinity was finally mathematized through set theory in the first years of the 20th century and the discussions on its potential or actual nature almost vanished. But, as we shall see in what follows, things still remain to be said on this issue.

During the last decades of the 19th century, Bolzano, Dedekind and most notably Cantor, inaugurated a new infinitist era in the history of mathematics, which included the birth of set theory. As expected, bijections and ellipses played a capital role in the foundation and subsequent development of the new infinitist theory. Interestingly, set theory was founded on a violation, the violation of the old Euclidian axiom of the whole and the part. Dedekind's foundational definition states that a set is infinite if it can be put into one-to-one correspondence with one of its proper subsets. For this very reason, Bolzano did not dare to complete and perfect the violation, a task that Dedekind and Cantor eventually rounded off.

A. León-Sánchez (✉)
I.E.S. Francisco Salinas, Salamanca, Salamanca, Spain
e-mail: aleon@interciencia.es

A. C. León-Mejía
International University of La Rioja, Logroño, Spain
e-mail: aleon@unir.net



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_11

The success of set theory as the fundamental theory of modern mathematics catapulted set theoretical infinitism to a hegemonic position. Yet, the controversy surrounding the infinite is not over. Before addressing the heart of this controversy, we need to examine the mathematical foundations of contemporary infinitism critically, and this will be our starting point.

It is somewhat ironic that set theory, the infinitist theory *par excellence*, contains mathematical instruments that may be used to call into question the formal consistency of the actual infinity hypothesis (i.e. the hypothesis of the existence of infinite collections as complete totalities). One of those instruments is ω, the first transfinite ordinal, the smallest ordinal greater than all finite ordinals. This first transfinite ordinal defines a type of well-order, called $\omega$-order, that characterizes the most basic infinite objects in transfinite mathematics such as ω-ordered sets and ω-ordered sequences. Most supertasks, for instance, are ω-ordered sequences of actions carried out in a finite time interval. Inevitably, ω-order implies a colossal asymmetry that is largely ignored in infinitist literature. In turn, this asymmetry gives way to a dichotomy that ultimately results in contradictions, whose ultimate cause can only be the axiom of infinity itself, which legitimates both ω and ω-order. In this paper, we shall resort to a formal version of Zeno dichotomy in order to examine this transition from asymmetry to inconsistency via the aforementioned dichotomy.

Two seminal papers published at the beginning of the second half of the 20th century laid the foundations for a new infinitist theory, independent of set theory, which has known subsequent developments throughout the last decades of the 20th century and the first years of the 21st, namely Thomson's work on what he called "super-tasks" (Thomson 1954–1955), and the criticism of this work by Paul Benacerraf (Benacerraf 1962). The success of Benacerraf's criticism somehow motivated the subsequent development of the new infinitist theory, i.e. of supertask theory. As in the previous case, the controversy is not yet over. Indeed, supertasks can also be used to question the hypothesis of actual infinity that is subsumed in the axiom of infinity. The objective of the present paper is to contribute to this criticism. As we shall see later on, we propose to start off exactly from where Benacerraf's arguments ended.

Supertasks are carried out by theoretical artefacts usually known as supermachines or infinite machines. The problem with machines, including theoretical supermachines, is not the (finite or infinite) number of actions they must carry out, but the machine's changes of state that are involved in each completed task. As is well known, the problem of change, another pre-Socratic inheritance, does not have a consistent solution within the space-time continuum. Therefore, anyone concerned with machines that undergo changes of state must, somewhat inconveniently, face an additonal problem: the problem of change. We shall see later on what may be done regarding this issue.

Definitions, procedures and proofs with infinitely successive steps are common fare in mathematics. Even though mathematics is not concerned with the way in which these infinitely successive steps could be carried out, the definitions, procedures and proofs of infinitely successive steps can be timetabled and converted into mathematical supertasks. These supertasks have the advantage of not requiring

the use of supermachines, theoretical as they may be. Hence it is possible to argue exclusively in mathematical terms, free of such theoretical complications, and to analyze the consequences one faces when assuming the axiom of infinity. Some of these mathematical supertasks will be discussed here.

Although supertasks are also discussed from a physical perspective, our goal here is not to involve physics in supertask theory but to illustrate the way supertasks could be used to call into question the hypothesis of actual infinity. By the same token, we shall also explain why such questioning is of great interest to experimental sciences such as physics. As a result, we shall only focus on conceptual supertasks. Furthermore, all of our arguments will be developed in a conceptual scenario entirely favorable to the actual infinity hypothesis, without any physical or chemical restriction limiting the discussion. Notwithstanding, we shall also take into account physics and infinity, particularly the restrictions that the Planck scale and Planck universal constants impose on supertasks and, what is more interesting, on the infinitist continuums involved in the special theory of relativity.

Platonism is the natural home of infinity and transfinite mathematics. In general, modern mathematics is essentially platonistic and a significant number of contemporary mathematicians are, at heart, platonists, which might be surprising from the perspective of the natural sciences. For this very reason, we shall conclude this essay by questioning platonism (platonic idealism) from a biological perspective, because we believe that evolutionary biology and neuroscience could shed some light on the classical conception of human knowledge embedded in platonism.

## 11.2 The Grounds of Transfinite Mathematics

As we already pointed out, Dedekind's foundational definition states that a set is infinite if it can be put into one-to-one correspondence with one of its proper subsets. It is, therefore, an operational definition of infinite sets based on the violation of Euclid's axiom of the whole and the part. Note that this definition says nothing about the potential or actual nature of the infinity involved. It is simply taken for granted that the infinity is actual infinity. In other words, it is presupposed that infinite sets are complete totalities. This is due to the fact that potentially infinite sets are not even considered in most mathematical discussions. Bolzano, Dedekind and Cantor unsuccessfully tried to prove the existence of actually infinite sets. Bolzano's proof is as follows:

> One truth is the proposition that Plato was Greek. Call this $p_1$. But then there is another truth $p_2$, namely the proposition that $p_1$ is true, [there is then another truth $p_3$, namely the proposition that $p_2$ is true]. And so ad infinitum. Thus the set of truths is infinite.
>
> Moore (2001: 112)

But this endless process ($p_1$ is true, then $p_2$ is true, then $p_3$ is true, then…) does by no means prove the existence of a final result as a complete totality. Dedekind's

proof is similar, and Cantor's one is even less successful: "Each potential infinite presupposes an actual infinity" (Hallet 1984: 25).

It is clear why the existence of an actually infinite set had to be established in axiomatic terms. This is precisely the objective of the axiom of infinity. In symbols:

$$\exists N(\varnothing \in N \wedge \forall x \in N(x \cup \{x\} \in N)).$$

Notice again that the axiom of infinity makes no reference to the type of infinity of the infinite set whose existence is being asserted. As in the case of Dedekind's definition, it is simply assumed that we are talking about the actual infinity.

Not only do one-to-one correspondences (bijections or exhaustive injections) play a role in Dedekind's foundational definition, they have also been used in a great variety of arguments throughout the history of the concept of infinity, mostly to prove (or disprove) the actual infinity hypothesis. They have been and continue to be (along with the inevitable short cuts) an essential instrument in the development of transfinite mathematics. We shall analyze them at the most basic foundational level of set theory.

It is sensible to assume that two sets $A$ and $B$ have the same number of elements if it is possible to pair each element of $A$ with an element of $B$ so that all the elements of $A$ and $B$ end up being paired (exhaustive injection). But it is also sensible to assume, for the very same reason, that if one or more elements of $B$ end up not being paired (non-exhaustive injection), then $A$ and $B$ do not have the same number of elements. The existence of both exhaustive and non-exhaustive injections between two infinite sets could indicate that they both have and do not have the same cardinality. Thus, the arbitrary distinction of the exhaustive injections to the detriment of the non-exhaustive ones could be concealing a fundamental contradiction in set theory. We shall begin by analyzing this "apparent" conflict.

If the notion of set is primitive (as it seems to be in the platonic scenario), we need operational definitions in which the pairing method plays a basic foundational role. Moreover, if sets have different sizes (cardinalities), we should establish an appropriate method for comparing them. We need to do so before considering the types of sets that may be defined according to their cardinalities, and before carrying out any other arithmetic or set theoretical operation. Exhaustive and non-exhaustive injections are the only known basic instruments fit to this purpose. It follows from this that the question whether the pairing method is appropriate to compare the cardinality of any two sets must be addressed at this very basic and foundational level of set theory. If the method turns out to be appropriate, we shall need to explain why non-exhaustive injections must be rejected, since the rejection could be pointing to a fundamental contradiction in set theory, i.e. the contradiction to the effect that infinite sets both have and do not have the same cardinality as some of their proper subsets.

It could be argued that infinite sets are defined as those that may be put into one-to-one correspondence with one of their proper subsets and that, for this very reason, it is possible to define exhaustive and non-exhaustive injections between any infinite set and some of its proper subsets. However, a simple definition does not guarantee that the defined object is consistent. Definitions may be inconsistent as well. Furthermore, the existence of a one-to-one correspondence between two

infinite sets isn't sufficient to show that they are actually infinite since both could, as a matter of fact, be potentially infinite. In the latter case, the infinite sets would not be regarded as complete totalities and so we would be pairing the elements of two incomplete totalities with the same cardinality. More importantly, under such conditions, the axiom of the whole and the part would not be violated.

Dedekind's definition could be based on one of the terms of the contradiction, say on the existence of an exhaustive injection between some infinite set and one of its proper subsets. The existence of a non-exhaustive injection between some infinite set and the same proper subset would constitute the other term of the contradiction. No one has ever explained, safe in circular terms, why having an exhaustive injection and a non-exhaustive injection with the same proper subset is not contradictory. The problem has simply been ignored or, better, justified in behalf of certain properties of infinite cardinals, all of them derived from the foundational definition that is being justified. We believe, however, that this problem needs to be addressed *before* we may define what an infinite set could be. Otherwise, set theory would lack a consistent basis.

The arithmetic peculiarities of transfinite cardinals, such as $\aleph_o = \aleph_o + \aleph_o$ and the like, could be used to explain why it is possible to define exhaustive and non-exhaustive injections between a set and one of its proper subsets. However, these arithmetic peculiarities are formal consequences of the assumption that the existence of sets may be put into exhaustive and non-exhaustive injections with some of their proper subsets. We cannot, therefore, on pain of unacceptable circular reasoning, resort to such arithmetic peculiarities to justify the existence of exhaustive and non-exhaustive injections between a set and one of its proper subsets. In short, at this foundational level of set theory, we cannot use posterior attributes of infinites sets derived from the foundational assumptions to justify the very same foundational assumptions.

If exhaustive and non-exhaustive injections were equally valid as instruments with which to compare the cardinality of any two sets, the actually infinite sets would be inconsistent. If, as a matter fact, they are not equally valid, we should explain in non-circular terms why exhaustive injections are valid instruments with which to compare the cardinality of infinite sets while non-exhaustive injections are not. Recall that both types of correspondences resort to exactly the same pairing method. If no (circular) reason may be given, we would have to admit that it is as legitimate to deem both types of injections valid instruments with which to compare cardinalities as it is legitimate not to. Leaving such a problem unsolved compels us to declare that the arbitrary superiority of exhaustive injections is a new axiom for the foundations of set theory.

Meanwhile, the foundations of set theory may rest on a contradiction. The paradoxes of reflexivity (like Galileo's celebrated paradox) are mere consequences of the assumption that there are both exhaustive and non-exhaustive injections between a set and one of its proper subsets. In other words, they are consequences of the violation of the old Euclidian axiom of the whole and the part. Clearly, they could also be reinterpreted as contradictions derived from the inconsistent nature of actually infinite sets (and, thus, as consequences of the axiom of infinity). But this alternative, as legitimate as it may be, has always been ignored.

The paradoxes of reflexivity are not the only paradoxes related to infinite sets. Burali-Forti's paradox of the set of all ordinals and Cantor's paradox of the set of all cardinals are other well-known examples. In such cases, we have, stricto sensu, genuine contradictions rather than paradoxes. According to Cantor, the inconsistent nature of these sets would be a consequence of their excessive infinitude, too close to the absolute infinity, the mother of all infinities that directly leads to God. It can be proved, however, that Cantor's inconsistency can easily be extended with the aid of Cantor's theorem of the power set. That extension proves (in naïve set theory) that each (finite or infinite) set of cardinal C originates no fewer than $2^C$ inconsistent sets, all of them infinite.

One may offer a summary of the proof as follows:

In naïve (non-axiomatic) set theory, the elements of a set may be sets, sets of sets, sets of sets of sets, etc. It therefore makes perfectly good sense to define the following relation **R** between two sets A and B: a set A is **R**-related to a set B (symbolically A **R** B) if B contains at least one element which forms part of the definition of at least one element of A.

For instance, the sets: $\{\{\{a, \{b\}\}\}, \{p\}, d, \{\{\{e\}\}\}, f\}$ and $\{a, b, c\}$ are **R**-related through the elements $a$ and $b$, while the sets $\{\{\{a, \{b\}\}\}, \{c\}, d, \{\{\{e\}\}\}, f\}$ and $\{1, 2, 3\}$ are not **R**-related.

Under such conditions, let X be any non empty set, and Y any of its subsets, and let us define the set $C_Y$ of all sets A that are not **R**-related to any set B that contains elements of Y, in the following way:

$$C_Y = \{A| \neg \exists B(B \cap Y \neq \emptyset \wedge A\,\mathbf{R}\,B)\}.$$

If $P(C_Y)$ is the power set of $C_Y$, then every element of $P(C_Y)$ is a subset of $C_Y$ and, consequently, a set of sets that are not **R**-related to any set that contains elements of Y:

$$\forall D \in P(_Y):\neg \exists B(B \cap Y \neq \emptyset \wedge D\,\mathbf{R}\,B).$$

Thus:

$$\forall D \in P(C_Y) :\ D \in C_Y.$$

It follows that the cardinality of $P(C_Y)$ is equal to, or less than, the cardinality of $C_Y$, which contradicts Cantor's theorem of the power set (the cardinality of any set is less than the cardinality of its power set).

Had we known the existence of such an infinity of inconsistent sets (far less infinite than either Cantor or Burali-Forti sets), transfinite set theory might have had a very different reception. This, however, hasn't been the case and for more than half a century all efforts have been directed towards the establishment of a foundation for a set theory free of inconsistencies. The goal was finally accomplished with the aid of a considerable number of ad hoc axioms. All of them, grouped in different ways, have contributed to the establishment of at least half a dozen axiomatic set theories.

We would indeed need several hundred pages to explain all these axiomatic restrictions. One may wonder if this is the best way of founding a formal science. The alternative to dealing with all these axioms is to consider a simpler explanation for all the inconsistencies arising from set theory, namely that actual infinity may after all be inconsistent, an avenue that needs to be explored.

## 11.3 ω-Order: From Asymmetry to Inconsistency

Cantor's *Beiträge* (*Cantor* 1955) was both the last and the most mathematical of his publications on transfinite arithmetic. In §6 we can read:

> The first example of a transfinite aggregate is given by the totality of finite cardinal numbers ν; we call its cardinal number "Aleph-zero" and denote it by "$\aleph_0$."

Cantor took the existence of that set, conceived as an actually infinite complete totality, for granted. It is reasonable to suppose that his platonistic and theological convictions prevented him from considering the existence of that complete totality as an initial foundational hypothesis (i.e. the axiom of infinity in contemporary set theories). In any case, he successfully derived from that infinite totality an infinity of growing transfinite cardinals and ordinals. In most of his proofs, Cantor made an extensive use of his concept of equivalent sets or equipotent sets, i.e. of sets that can be put into one-to-one correspondence. This concept also illustrates the great importance of bijections, and the role played by the violation of the axiom of whole and the part in the foundation of transfinite mathematics.

Theorem I (*Cantor op. cit.*: Part II, §14), which proves the existence of ordinals as limits of increasing fundamental sequences of ordinals, is particularly revealing in this respect. In Cantor's terminology, these are the ordinals of the second class, or of the second kind. The first transfinite ordinal, ω, is the first of the second class, and of the second kind of transfinite ordinals: it is the smallest of all ordinals greater than all finite ordinals. This ordinal defines a type of well order usually known as ω-order. The set of natural numbers in their natural order of precedence is a well-known example of an ω-ordered set. It is important to highlight that ω-order and ω-ordered sequences will play an important role in what follows. For now, let us note that their existence is formally deduced from the axiom of infinity.

Let us begin our journey from ω-order to ω-inconsistency. The first leg of the journey will lead us from ω-order to ω-asymmetry. To begin with, let us consider any ω-ordered sequence $a_1, a_2, a_3, \ldots$ In these types of sequences, there is a first element $a_1$, and each element $a_n$ has an immediate predecessor $a_{n-1}$ (except $a_1$) as well as an immediate successor $a_{n+1}$, so that there is no last element. As a consequence, every element in the sequence has a finite number of predecessors and an infinite number of successors. We call this kind asymmetry "ω-asymmetry." Since infinitist mathematics takes ω-ordered sequences to be complete totalities, we could travel through each of the successive elements of the sequence and complete the journey in a finite time. But even if we managed to complete this journey, we would never reach an

element with an infinite number of predecessors and a finite number of successors. From the very start to the very end of this infinitist excursion, we would always be dealing with elements that have a finite number of predecessors and an infinite number of successors. This sort of Red Queen's race to nowhere is known as ω-asymmetry.

To grasp the colossal magnitude of ω-asymmetry, consider a finite straight-line segment AB a trillion times greater than the diameter of the visible universe ($9.3 \times 10^{10}$ light years). Consider a point C on AB, arbitrarily close to B, and let us assume that AB is ω-partitioned. Whatever the ω-partition may be, only a finite number of parts will lie within AC, while infinitely many of them will lie within CB, CB being a trillion times smaller than, e.g., Planck length ($1.62 \times 10^{-33}$ cm) which, in turn, is inconceivably smaller than, e.g., the smallest of the atomic nuclei. Due to ω-asymmetry, there is no way to devise a less asymmetric partition in case the partition were ω-ordered, i.e. in case it were the smallest of the infinite partitions. Moreover, whatever part you may consider (even within CB), it will always have a finite number of preceding parts and an infinite number of succeeding ones. This is how ω-asymmetry works.

Let us now travel from ω-asymmetry to ω-dichotomy. In order to do so, let us consider the X axis of the Euclidian space $R^3$. Let us assume that its interval (0,1) is partitioned by the sequence $\{z_n\}$ of points defined by:

$$Z_n = (2^n - 1)/2^n, \quad \forall\ n \in N$$

where N is the set of natural numbers.

For well-known historical reasons, the points $\{z_n\}$ will be referred to as Z-points (for "Zeno's points"). Now, consider a mass point P moving through the X axis from point 2 to point −2 at a finite and uniform velocity v. Assume that at instant t = 0, P is at point 1. At instant 1/v, it will be at point 0, which means that it has gone through all Z-points (since they form a complete totality). Let f(t) be the number of Z-points that P has gone through at instant t, for any t within the time interval [0, 1/v]. As a consequence of ω-asymmetry we will have:

$$\begin{array}{ll} f(t) = 0 & \text{if } t = 0 \\ f(t) = \aleph_o & \text{if } t > 0 \end{array}.$$

There is no instant t in [0, 1/v] at which f(t) = n, n being any natural number greater than zero. Otherwise, we would have to deal with the existence of the last n elements of an ω-ordered sequence, something that is indeed impossible in ω-order (ω-asymmetry). Keep in mind that f is well defined for every t within the interval [0, 1/v]. It maps the set [0, 1/v] onto the set $\{0, \aleph_o\}$. In other words, f defines a dichotomy. So, with respect to the number of Z-points P has gone through, P can only exhibit two states: the state P(0), at which it has gone through 0 Z-points, and the state $P(\aleph_o)$ at which it has gone through $\aleph_o$ Z-points. There are no intermediate states P(n) at which P would have gone through a finite number n of Z-points. P will always be either at P(0) or at $P(\aleph_o)$. This is ω-dichotomy, a consequence of ω-asymmetry which is itself a consequence of ω-order which is, in turn, formally derived from the axiom of infinity.

Finally, let us travel from ω-dichotomy to ω-inconsistency. First, notice that the points of the interval (0, 1) are densely ordered (there are infinitely many other different points between any two of them), whereas Z-points are not. Each Z-point has an immediate predecessor (except the first one) and an immediate successor, and no other Z-point exists between any two successive Z-points $z_n$, $z_{n+1}$. In addition, there is always a distance $d_n = z_{n+1} - z_n = 1/2^{(n+1)}$ greater than zero between any two successive Z-points $z_n$, $z_{n+1}$ (ω-separation). Consequently, at any finite velocity, P may go through them only successively, i.e. one by one, one at a time, one after the other, and in such a way that it takes a time greater than zero to go from any Z-point to its immediate successor or to its immediate predecessor.

We now know that P travels from point 1 to point 0 at a finite and uniform velocity v, so it must become $P(\aleph_o)$ from P(0) as it travels from point 1 to point 0. As a consequence of ω-order, there is no last Z-point to begin the transition P $(0) \rightarrow P(\aleph_o)$. Thus, it will be impossible for us to calculate, either the distance P must travel to become $P(\aleph_o)$ when starting from P(0), or the time it takes P to complete the transition. But the transition takes place anyway, even if we are unable to describe the way in which it does, since P(0) = 0 and $P(1/v) = \aleph_o$.

We shall now prove that the transition $P(0) \rightarrow P(\aleph_o)$ must be instantaneous. For this purpose, let t be any real number greater than zero, and let us assume that P makes the transition $P(0) \rightarrow P(\aleph_o)$ at time t. Let t′ be any element in the interval (0, t). According to ω-dichotomy, and since $t' > 0$, we will have $P(t') = \aleph_o$. Therefore, at t′, the transition $P(0) \rightarrow P(\aleph_o)$ has already been completed. Consequently, that transition lasts a time less than t, which is any real number greater than zero. We must conclude that the transition $P(0) \rightarrow P(\aleph_o)$ lasts a time less than any real number greater than zero, in other words that the transition $P(0) \rightarrow P(\aleph_o)$ lasts a null time: it can only be instantaneous. It is worth noting that we are not facing a problem of indeterminacy due to the fact that we cannot measure the duration of the transition, but with an impossibility directly derived from ω-dichotomy: the transition $P(0) \rightarrow P(\aleph_o)$ lasts a time less than any real number greater than zero, and this is possible only in case it takes a null time to achieve the transition.

We have just proved that the transition $P(0) \rightarrow P(\aleph_o)$ must be instantaneous. This implies that P must go through $\aleph_o$ successive Z-points instantaneously at a finite velocity v. This is nevertheless impossible because there is a distance greater than zero between any two successive Z-points (ω-separation). Travelling a distance greater than zero at a finite velocity always takes a time greater than zero. It is therefore impossible for P to go through any two successive Z-points instantaneously. At its finite velocity v, P cannot go through several Z-points simultaneously. It can only go through them successively, one after the other, so that a time greater than zero always elapses between going through one Z-point and going through the next. Oddly enough, we must conclude that the transition $P(0) \rightarrow P(\aleph_o)$ must be instantaneous even though it may not be. In short: the transition $P(0) \rightarrow P(\aleph_o)$ takes place; due to ω-dichotomy, $P(0) \rightarrow P(\aleph_o)$ can only be instantaneous; due to ω-separation, $P(0) \rightarrow P(\aleph_o)$ cannot be instantaneous at a finite velocity.

In addition to Z-points, we could also consider Z*-points $\{z_i^*\}$ defined within (0, 1) by:

$$Z_n^* = 1/2^n, \ \forall n \ \in N.$$

f(t) being the number of Z*-points that P must go through at any instant t within [0, 1/v], f defines a new ω-dichotomy: with respect to the number of Z*-points P must go through, P can only exhibit two states: the state $P(\aleph_0)$, in case there still are $\aleph_o$ points for P to go through, and the state P(0) at which no Z*-point remains. Intermediate states P(n) at which P would have to go through only a finite number of Z*-points are impossible. An argument similar to the argument about Z-points leads to the following conclusion concerning the $P(\aleph_o) \rightarrow P(0)$ transition: the transition $P(\aleph_o) \rightarrow P(0)$ takes place; due to ω-dichotomy, $P(\aleph_o) \rightarrow P(0)$ can only be instantaneous; due to ω-separation, $P(\aleph_o) \rightarrow P(0)$ cannot be instantaneous at a finite velocity.

This is ω-inconsistency, an (almost) direct consequence of ω-dichotomy which, in turn, is a formal consequence of ω-asymmetry (the existence of complete totalities in which each and every element has a finite number of predecessors and an infinite number of successors), which, in turn, is an immediate consequence of the ω-order derived from ω. Recall that ω is the least transfinite ordinal of the second class, and of the second kind, whose existence Cantor deduced from the assumption of the existence of a complete infinite totality of finite cardinals (i.e., in modern terms, the axiom of infinity).

Just as in the case of transitions $P(0) \rightarrow P(\aleph_o)$ and $P(\aleph_o) \rightarrow P(0)$, ω-inconsistency will also come forth in any ω-ordered sequence of actions $\{a_i\}$ successively carried out at each of the successive instants of $\{t_i\}$. This means that an infinite number of these actions would have to be carried out instantaneously, while they can only be carried out successively, so that a time $\Delta_n t = t_{n+1} - t_n$ greater than zero always elapses between any two of these successive actions. Deriving a contradiction from an axiom should be a sufficient reason to seriously consider the possibility that the axiom might after all be inconsistent. Nevertheless, for mysterious reasons, this seems to be still not enough when that axiom happens to be the axiom of infinity. Let us, therefore, continue to examine the issue.

## 11.4 Benacerraf and Thomson: A Seminal Discussion

The concept of supertask was already implicit in many classical discussions involving infinity, e.g. in Zeno's dichotomy and in Aristotle's criticism of it. In the 14th century, Gregory of Remini explained how infinitely many successive actions could be carried out in a finite time in the following way:

> If God can endlessly add a cubic foot to a stone — which He can — then He can create an infinitely big stone. For He need only add one cubic foot at some time, another half an hour later, another a quarter of an hour later than that, and so on ad infinitum. He would then have before Him an infinite stone at the end of the hour.
>
> Moore (2001: 53)

Disputation about such claims became popular, at least in an academic setting, at the beginning of the 1950s. Black's machines meant to prove the impossibility of performing an infinite number of successive actions (Black 1950–1951). Black's arguments were then discussed by Taylor (Taylor 1951) and Watling (Watling 1952). Thomson's paper was precisely motivated by these discussions (Thomson 1954–1955). Thomson introduced the term "super-task" and developed several arguments, among which the famous lamp argument (which we shall discuss in the next section), purporting to prove their impossibility. Benacerraf successfully criticized Thomson's arguments in a seminal paper published in 1962, and this somehow gave rise to the birth of a new infinitist theory in the last decades of the 20th century, i.e. supertask theory (Benacerraf 1962).

Most supertasks are ω-supertasks, i.e. ω-ordered sequences of successive actions (tasks) carried out at the successive instants of a strictly increasing ω-ordered sequence of instants within a finite interval of time. We shall restrict our attention to conceptual ω-supertasks. We shall assume that they are all carried out along the same strictly increasing and ω-ordered sequence $\{t_i\}$ of instants within the same finite interval of time $[t_a, t_b]$, each action $a_i$ being carried out at the precise instant $t_i$, and $t_b$ being the mathematical limit of the sequence $\{t_i\}$. They will be denoted by "$\{a_i, t_i\}$," "$\{b_i, t_i\}$," "$\{c_i, t_i\}$," etc.

The possibility of carrying out an uncountable infinity of successive actions has been examined and ruled out by Clark and Read (Clark and Read 1984). Their proof was based on Cantor's proof of the impossibility of uncountable partitions of the real line (Cantor 1885). Let us note that Cantor's proof is not an independent proof but an immediate corollary of his theorem on the countable nature of the set of rational numbers (more on this below). Supertasks have also been examined from the perspective of non-standard analysis. But, as far as we know, the possibility of carrying out hypertasks along hyperreal intervals of time has not been discussed, despite the fact that finite hyperreal intervals may also be divided into hypercountably many successive infinitesimal intervals (hyperfinite partitions).

Only conceptual ω-supertasks will be considered here. So let us begin with the supertask achieved by Thomson's reading lamp, i.e. "[a] lamp[] [that] has a button in the base. If the lamp is off and you press the button, the lamp goes on, and if the lamp is on and you press the button, the lamp goes off" (Thomson 1954–1955: 5). Let us add the following constraints to Thomson's description: (a) the lamp has only two states: on and off; (b) the only way to change the state of the lamp is by pressing the button; (c) each change of state takes place at a precise and definite instant; (d) the pressing of the button and the corresponding lamp's change of state are instantaneous and simultaneous events.

Most variants of Thomson's lamp have been proposed with the *physical* possibility of performing Thomson's supertask in mind. We shall take our Thomson's lamp to be a theoretical device whose role is to contribute to our examination of the formal consistency of the axiom of infinity.

The problem of change surfaces in all discussions of supertasks carried out by supermachines undergoing changes of state. As we know, any canonical change from state A to state B (without intermediate states) poses a problem that remains

unresolved in the space-time continuum. It has been claimed for a long time that canonical changes could be inconsistent. As a matter of fact, if the change has a duration greater than zero, the changing object can only be in an unknown state (different from both A and B) when the change takes place. The problem of change thus arises in terms of a *new* change between the state A and that still unknown state, and so on and so forth. On the other hand, if the change is instantaneous, it cannot take place in the space-time continuum since, in this continuum, *no* instant has an immediate successor and a time greater than zero *always* elapses between any two of its instants. Things could be quite different in discrete spacetimes, where immediate successiveness is an essential characteristic of both space and time. But let us focus for now on what happens in the space-time continuum.

We shall ignore the problem of change for the sake of discussion and assume that change is instantaneous so as to discuss supertasks from the perspective of the axiom of infinity, focusing our attention on ω-ordering and on the corresponding ω-asymmetries, ω-dichotomies and ω-inconsistencies. Bear in mind that, from a purely theoretical point of view, ω-dichotomies and ω-inconsistencies have nothing to do with the problem of change, except for the fact that infinite sequences are complete totalities.

Let $\{c_i, t_i\}$ be Thomson's supertask and let us assume that each click $c_i$ is performed at the precise instant $t_i$ of the strictly increasing sequence $\{t_i\}$ of instants within the finite interval $(t_a, t_b)$, $t_b$ being the limit of $\{t_i\}$.

Thomson claims that:

> [The lamp] cannot be on, because I did not ever turn it on without at once turning it off. It cannot be off, because I did in the first place turn it on, and thereafter I never turned it off without at once turning it on. But the lamp must be either on or off. This is a contradiction.
>
> Thomson (1954–1955: 5)

Benacerraf objects that:

> The only reasons Thomson gives for supposing that his lamp will not be off at $t_b$ are ones which hold only for times before $t_b$. The explanation is quite simply that Thomson's instructions do not cover the state of the lamp at $t_b$, although they do tell us what will be its state at every instant between $t_a$ and $t_b$ (including $t_a$). Certainly, the lamp must be on or off (provided that it hasn't gone up in a metaphysical puff of smoke in the interval), but nothing we are told implies which it is to be. The arguments to the effect that it can't be either just have no bearing on the case. To suppose that they do is to suppose that a description of the physical state of the lamp at $t_b$ (with respect to the property of being on or off) is a logical consequence of a description of its state (with respect to the same property) at times prior to $t_b$.*
>
> * "$t_a$" and "$t_b$" appear respectively as "$t_0$" and "$t_1$" in Benacerraf's paper.
>
> Benacerraf (1962: 768)

We agree with Benacerraf that we cannot deduce the state of the lamp at $t_b$ from the sequence of its previous changes. We also assume that certain properties of the sequence of changes that hold while the number of changes is finite may be not satisfied in case that number is infinite. But, as we shall see, the discussion cannot end here. Benacerraf's conclusion that "the lamp must be on or off […], but nothing

we are told implies which it is to be" will be the starting point of our extension of Benacerraf's argument about $\{c_i, t_i\}$.

We may not infer that a machine is not the same as it was before and while performing some supertask from our ignorance of what the state of the machine is once it has indeed achieved the supertask. The machine could only change its (theoretical or physical) nature after completing the supertask, otherwise it would be impossible for the machine to complete it. There is no reason to assume that, if we define a theoretical machine as one which is to achieve a conceptual supertask, that machine is no longer the one it was defined to be once it has achieved the supertask, no matter what its current state might be (except arbitrarily, for the sake of convenience). In other words, ignoring the state of the machine isn't quite the same thing as ignoring the nature of the machine (i.e., in our conceptual scenario, its formal definition).

Consequently, we shall assume that, once it has achieved a supertask, the conceptual objects that participated in it (regardless of their current states) continue to be the same objects they were before and during the carrying out of the supertask. If, e.g., x is a rational variable and we redefine it a (finite or infinite) number of successive times, we still take x to be a rational variable once such redefinitions have been provided, and not, say, a red hat or a neutron star. By the same token, if T is a table of real numbers within (0, 1) whose rows are permuted any finite or infinite number of times, once the permutation has occurred, T will continue to be a table with the very same real numbers it had before its rows were permuted. So, unnecessary as it may seem, we shall begin by assuming the following hypothesis:

H0: The definition of a conceptual object does not change as a consequence of any finite or infinite sequence of successive actions being carried out with the help of that object.

More specifically, we shall assume that formal definitions, laws, conditions and constraints are never arbitrarily violated as a consequence of having carried out a finite or infinite number of actions. Denying H0 would have destructive consequences for transfinite mathematics. For instance, after providing a recursive definition (or procedure, or proof) of infinitely many successive steps (that could also be scheduled in the form of a supertask), we couldn't assert anything about the defined object. And this would also happen to all the axioms, definitions and theorems involved.

If nothing could be asserted of a conceptual object once a supertask has been achieved, nothing could be asserted either of any mathematical object or result obtained through a sequence of infinitely many successive steps. Obviously, under such conditions, transfinite mathematics would remain empty of content (more on mathematical supertasks below). As we shall see, we could also define conditional supertasks (mathematical supertasks, in particular) in such a way that each task would be achieved only in case certain conditions are satisfied. In this case, it would indeed be impossible for us to know whether the number of achieved tasks is finite or infinite. Even if we admit that the state of a conceptual object cannot be deduced from its previous states while performing a supertask, H0 guarantees that its formal definition does *not* change as a consequence of such a carrying out.

## 11.5 Thomson's Lamp Revisited

According to H0, Thomson's lamp will remain the same Thomson lamp before, during and after the completion of supertask $\{c_i, t_i\}$. Consequently, at $t_b$, after the completion of $\{c_i, t_i\}$, the lamp will be in a certain state $S_b$. We are not interested in knowing whether the lamp is on or off at $S_b$, although by definition a Thomson lamp can only be either on or off. Some authors have claimed that it could be in some exotic state different from these two states. We must nevertheless insist that if the lamp can be in some exotic state other than either on or off, then it isn't by definition, a Thomson lamp.

We know that the state of the lamp is $S_b$ at $t_b$ and we are able to prove that at any instant prior to $t_b$ it is impossible for the lamp to have reached $S_b$, whatever that state might be. Let t be any instant prior to $t_b$. Since $t_b$ is the limit of the sequence $\{t_i\}$, there will be a $t_v$ in $\{t_i\}$ such that $t_v \leq t < t_{v+1}$, which means that only a finite number v of clicks have been carried out at t (ω-asymmetry). As a consequence, $S_b$ cannot originate at t for any t within $(t_a, t_b)$. Therefore, $S_b$ being the state of the lamp at $t_b$, the state $S_b$ can only originate at $t_b$ and at no other instant. Notice that this conclusion is a direct consequence of the fact that $t_b$ is the mathematical limit of $\{t_i\}$ together with the assumption that $\{c_i, t_i\}$ has been carried out along the successive instants of the strictly increasing sequence $\{t_i\}$.

Notice also that while $t_b$ is the mathematical limit of the strictly increasing and upper bounded sequence of real numbers (successive instants) $\{t_i\}$, the state $S_b$ is not the mathematical limit of the sequence of states $\{S_i\}$ = on/off/on/, off/on/off… that the lamp undergoes as a consequence of $\{c_i, t_i\}$. Recall that oscillating sequences do *not* have a limit. Therefore, $S_b$ is the state of a Thomson lamp that originates at some specific instant $t_b$, otherwise the supertask would not have been completed. This is indeed all that may be said about $S_b$ and the supertask $\{c_i, t_i\}$.

According to the definition given above, a Thomson lamp only changes its state when the button is clicked. So we cannot claim that the lamp may change its state in virtue of yet unknown reasons. Remember that, according to H0, a lamp that changes its state for unknown reasons is *not*, by definition, a Thomson lamp. Since $t_b$ is the specific and definite instant at which $S_b$ originates, the button of the lamp had to be clicked at $t_b$ (the clicking and the corresponding change of state are instantaneous and simultaneous events that take place at a specific and definite instant). Yet this is impossible because at $t_b$ the supertask $\{c_i, t_i\}$ has already been achieved. It follows that $t_b$ is the first instant after performing $\{c_i, t_i\}$ and that the button of the lamp has not yet been clicked at $t_b$.

Let f(t) be the number of clicks to be performed at t within the closed interval $[t_a, t_b]$. As a consequence of ω-order and ω-asymmetry, each $c_i$ of $\{c_i\}$ has infinitely many successors. Consequently, we shall have:

$$f(t) = \aleph_o \quad \text{if } t < t_b$$
$$f(t) = 0 \quad \text{if } t = t_b$$

which means that, for any natural number n, there is no instant t at which f(t) = n. Otherwise, the (impossible) last n elements of an ω-ordered sequence would exist; in other words, there would exist an element of the sequence with a finite number n of successors. Therefore, with respect to the number of clicks to occur, a Thomson lamp may only have two states:

TL($\aleph_0$), at which $\aleph_0$ clicks still have to occur

or

TL(0), at which no further clicks may occur.

A similar argument to the one provided above for Z*-points proves that the transition TL($\aleph_0$) → TL(0) can only be instantaneous, and hence that $\aleph_0$ clicks have to occur simultaneously. This contradicts the fact that the button of the lamp is clicked successively and that each click $c_i$ takes place at instant $t_i$ in such a way that a time $\Delta_i t = t_{i+1} - t_i$ greater than zero always elapses between any two successive clicks $c_i$ and $c_{i+1}$ (ω-separation).

In order to illustrate the difficulty brought about by the ω-dichotomy of $\{c_i, t_i\}$, consider a box BX containing a denumerable sequence $\{b_i\}$ of labelled balls $b_1$, $b_2$, $b_3$, … and assume that we remove the balls from the box one by one in such a way that at each click $c_i$ (i.e. at each instant $t_i$), we remove the ball $b_i$ from the box. At $t_b$ all the balls will have been removed from BX, exactly as the one-to-one correspondence $g(t_i) = b_i$ proves. If f(t) is the number of balls to be removed at instant t, we obtain the same ω-dichotomy that we obtained in the case of Thomson's lamp. So, despite the fact that all the balls are removed one by one, one after the other, and in such a way that a time $\Delta t = t_{i+1} - t_i > 0$ always elapses between the extractions of any two successive balls $b_i$, $b_{i+1}$ (ω-separation), the box BX will *never* contain a finite number of balls. BX is emptied by the successive removal of all of the balls *one by one*, but it will never contain …5, 4, 3, 2, 1, 0 balls. The number of balls inside the box will always be either $\aleph_0$ or 0; it will indeed suddenly change from $\aleph_0$ to 0.

Furthermore, as in the case of $S_b$ (and for the same reasons), the change can only be instantaneous. So an infinite number of balls would have to be removed from the box simultaneously, which is incompatible with the fact that all of them are removed successively, as the bijection $f(t_i) = b_i$ proves. All things considered, one may wonder whether we are facing a new infinitist extravagance or a mere inconsistency. Notice that we are not subtracting cardinals but removing balls from a box under the restriction of a dichotomy formally derived from ω-order, and thus from the axiom of infinity. The substraction of cardinals is indeed a suspicious transfinite arithmetic operation: it is admissible only in some specific cases (e.g. in the Tarski-Bernstein and the Tarski-Sierpinski theorems). Sometimes, it is utterly inconsistent (see, e.g. the Faticoni argument and cognates).

For strictly illustrative purposes, and without going into further details, we shall now formalize the Thomson-Benacerraf dispute.

Consider the following expressions and their corresponding symbolic counterparts:

- Thomson's lamp on at instant t: *[t]
- Thomson's lamp off at instant t: o[t]
- Thomson's lamp on along the interval ($t_a$, $t_b$): *($t_a$, $t_b$)
- Thomson's lamp off along the interval ($t_a$, $t_b$): o($t_a$, $t_b$)
- Click at t the lamp being previously on: c{[t], *}
- Click at t the lamp being previously off: c{[t], o}
- Click at least one time along the interval ($t_a$, $t_b$) the lamp being previously on: c {($t_a$, $t_b$), *}
- Click at least one time along the interval ($t_a$, $t_b$) the lamp being previously off: c {($t_a$, $t_b$), o}.

The word "previously" in "the lamp being previously on (off)," does matter a great deal here. Recall that in the spacetime continuum no instant has an immediate preceding (or succeeding) instant in the way that, e.g., the natural number 5 has an immediate predecessor (the natural number 4) and an immediate successor (the natural number 6). As noted above, this is why the problem of change remains unsolved in the spacetime continuum. The problem bears upon all changes we may think of, whether theoretical or experimental, so that we must indeed leave it on the side if we want to discuss fruitfully the question of the state changes of a Thomson lamp.

With the help of our symbolism, we can formalize some fundamental laws of Thomson's lamp, e.g. the following axioms (definition of the lamp):

$$
\begin{aligned}
&c\{[t], o\} \Rightarrow^{*} [t].\\
&c\{[t],^{*}\} \Rightarrow o[t].\\
&^{*}[t] \vee o[t]\\
&\neg(^{*}[t] \wedge o[t])
\end{aligned}
$$

and the following derived laws:

$$
\begin{aligned}
&c\{(t_a, t_b), o\} \Rightarrow \exists t \in (t_a, t_b) :^{*} [t]\\
&c\{(t_a, t_b),^{*}\} \Rightarrow \neg^{*}(t_a, t_b)\\
&o[t_b] \Rightarrow \neg^{*}[t_b, \infty)\\
&\text{etc.}
\end{aligned}
$$

We shall focus on the following two laws:

$$
\begin{aligned}
\text{BT1: } & c\{(-\infty, t_b),^{*}\} \wedge^{*} [t_b, \infty) \Rightarrow \exists\, t \leq t_b : c\{[t],o\} \wedge \neg\\
& \quad c\{(t, \infty),^{*}\}\\
\text{BT2: } & c\{(-\infty, t_b), o\} \wedge o[t_b, \infty) \Rightarrow \exists\, t \leq t_b : c\{[t],^{*}\} \wedge \neg\\
& \quad c\{(t, \infty), o\}.
\end{aligned}
$$

BT1 reads: if the lamp's button has been clicked at least once within the interval ($-\infty$, $t_b$), the lamp being previously on, and the lamp stays on from $t_b$, then there is

an instant t equal or prior to $t_b$ such that the button is clicked at t, the lamp being previously off, and the button is no longer clicked from t. BT2 reads similarly except that we must replace "on" with "off" and vice versa.

Let us now prove BT1 (the proof of BT2 may be provided along similar lines).

$$\text{H1: Assume that } \neg\exists t \leq t_b\text{: } c\{[t], o\}.$$

We will have:

$$\neg c\{(-\infty, t_b], o\}.$$

On the other hand, according to the antecedent of BT1 we have:

$$c\{(-\infty, t_b),^*\} \Rightarrow \exists t < t_b : c\{[t],^*\}$$

which means o[t].

From:

$$\neg\, c\{(-\infty, t_b], o\} \text{ and } o[t], \text{ being } t < t_b$$

we derive $o[t_b]$ and $\neg^*[t_b, \infty)$, which tells against the second term of the antecedent in BT1. Therefore, if that antecedent is true, H1 is false.

$$\text{H2: Assume that: } \neg\exists\, t \leq t_b : \neg c\,\{[t, \infty),^*\}.$$

We will have:

$$c\{[t_b, \infty),^*\}$$

which tells against the second term $^*[t_b, \infty)$ of the antecedent of BT1. So, if this antecedent is true, H2 must be false. The falsehood of H1 and H2 proves BT1. Notice that BT1 is not derived *à la* Thomson, from the successively performed clicks. BT1 is a law directly derived from the laws that define Thomson's lamp. Therefore, if we assume H0, BT1 must hold before, during and after the occurrence of any (finite or infinite) number of clicks.

Consider again the supertask $\{c_i, t_i\}$. Assume that the state $S_b$ is on (a similar argument can be developed if it were off, with BT2 in the place of BT1). Under such conditions, the antecedent of BT1 would be true. Therefore, its consequent would also be true. However, it is false. On the one hand, if $t < t_b -$ ($t_b$ being the limit of the sequence $\{t_i\}$), there would exist some $t_v$ in $\{t_i\}$ such that $t_v \leq t < t_{v+1}$, and hence only a finite number v of clicks would have occurred. On the other hand, t cannot be $t_b$, because at $t_b$ the button of the lamp has not yet been clicked. Consequently, t cannot be an element of $(\leftarrow, t_b]$. Therefore, the carrying out of superstask $\{c_i, t_i\}$ implies the violation of BT1, which flies in the face of H0.

## 11.6 On Marbles and Boxes

We sometimes speak of paradoxes when inconsistencies are at stake as, e.g., in the case of the Burali-Forti and Cantor paradoxes. Ross's so-called paradox might also belong to this category. It is indeed a supertask. Before considering the Ross-Littlewood paradox, we shall introduce expofactorial and n-expofactorial numbers in order to define two variants of Ross's supertask (expofactorials were also introduced independently by C. Pickover in Pickover 1995). Although we do not need such numbers to define supertasks, they make things easier for us and help illustrate how finite natural numbers may be so large as to prove offensive.

The expofactorial $n^!$ (note that the factorial symbol ! appears as an exponent) of a number n is the factorial n! raised to a power tower of order n! of the same exponent n! Thus the expofactorial of 2 is:

$$2^! = 2! \wedge 2! \wedge 2! = 2 \wedge 2 \wedge 2 = 2^4 = 16.$$

We cannot calculate the expofactorial of 3 ($3^!$):

$$\begin{aligned} 3^! &= 3! \wedge 3! \wedge 3! \wedge 3! \wedge 3! \wedge 3! \\ &= 6 \wedge 6 \wedge 6 \wedge 6 \wedge 6 \wedge 6 \\ &= 6 \wedge 6 \wedge 6 \wedge 6^{46656} \\ &= 6 \wedge 6 \wedge 6^{26591197721532267796824894043879185990534220026992430066043278949707355987388290912134229290...} \end{aligned}$$

where the incomplete exponent of the last expression of $3^!$ has no less than 36,306 digits (it is roughly ten pages long). We haven't been able to calculate the next step, not even with the help of a big integer supercalculator. And there still remain three steps to go. So $3^!$ is such a large number that we cannot calculate its precise sequence of numerals (it isn't a mere sequence of zeros), not even with the help of the most powerful computers. Imagine the case of $9^!$, let alone of $100^!$.

Expofactorials are very small compared to n-expofactorials, which are recursively defined from expofactorials: The 2-expofactorial of a number n (denoted "$n^{!2}$," is the expofactorial $n^!$ raised to a power tower of order $n^!$ of the same exponent $n^!$:

$$n^{!2} = n^! \wedge n^! \wedge \ldots (n^! \text{ times}) \ldots \wedge n^!.$$

The 3-expofactorial of a number n, denoted "$n^{!3}$," is the 2-expofactorial $n^{!2}$ raised to a power tower of order $n^{!2}$ of the same exponent $n^{!2}$:

$$n^{!3} = n^{!2} \wedge n^{!2} \wedge \ldots (n^{!2} \text{ times}) \ldots \wedge n^{!2}.$$

The 4-expofactorial of a number n, denoted "$n^{!4}$," is the 3-expofactorial $n^{!3}$ raised to a power tower of order $n^{!3}$ of the same exponent $n^{!3}$:

$$n^{!4} = n^{!3} \wedge n^{!3} \wedge \ldots (n^{!3} \text{ times}) \ldots \wedge n^{!3}$$

and so on and so forth. Three arithmetical symbols—"9," "$^{!}$," and "$^{9}$,"—suffice to denote a number (9-expofactorial of 9) so large that the standard writing of its precise sequence of figures would require a volume of paper trillions of times greater than the volume of the visible universe. As we noted above, they are so large that they prove offensive. Still, being finite, they are much smaller than the smallest of the infinite cardinals $\aleph_o$ (or than the improper real number $\infty$). In some of the following discussions we shall resort to the number $9^{!9}$ which, for the sake of simplicity, will be denoted by the letter H (for "huge").

Let us now consider the following supertask, our first variant of Ross's supertask: at each instant $t_i$ of $\{t_i\}$ we add H marbles (i.e. $9^{!9}$) to an initially empty box A. If the index i is an integer multiple of H (i.e. $t_H$, $t_{2H}$, $t_{3H}$,…), then one marble is added to another initially empty box B. At $t_b$, once the supertask has been completed, A and B will contain the same number of marbles, i.e. $\aleph_o$. From the transfinite arithmetic perspective, there is nothing remarkable in this conclusion because transfinite cardinals satisfy equations such as $\aleph_o = (\aleph_o)^H$ and the like. One might nevertheless feel frustrated from a conceptual or philosophical point of view.

On the one hand, $\aleph_o$ is the least transfinite number greater than all finite integers. In this respect, it is the upper limit of any strictly increasing ω-ordered sequence of natural numbers. Since the number of marbles in each box forms a strictly increasing ω-ordered sequence of natural numbers, at $t_b$ both boxes contains the same number $\aleph_o$ of marbles. This is fine, but, since the supertask progresses, there is a third strictly increasing and ω-ordered sequence of natural numbers $\{d_i\}$, namely the difference in the number of marbles in A and B:

$$\begin{aligned} \{d_i\} &= H, 2H, 3H, \ldots H^2 - 1, H^2 + H - 1, H^2 + 2H - 1, H^2 \\ &\quad + 3H - 1, \ldots H^3 - 2 \ldots \\ d_i &= H_i^{1+a} + b_i H - a_i \end{aligned}$$

where $a_i$ = Int(i/H) and $b_i$ = (i mod H).

So, if at $t_b$ the number of marbles in A is the limit $\aleph_o$ of the sequence

$$\{iH\} = H, 2H, 3H, \ldots$$

and the number of marbles in B at the same instant $t_b$ is the limit $\aleph_o$ of the sequence

$$\{i\} = 1, 2, 3, \ldots$$

we may wonder why the difference in the number of marbles in A and in B at $t_b$ is *not* the limit $\aleph_o$ of the sequence

$$\{H_i^{1+a} + b_i H - a_i\} = H, 2H, 3H, \ldots H^2 - 1, H^2 + H - 1, H^2 + 2H - 1, H^2 + 3H - 1, \ldots H^3 - 2 \ldots$$

How is it possible that the difference be 0 at $t_b$? Recall that at $t_b$ both boxes contain the same number of marbles. Notice also that we are discussing the limits of strictly increasing ω-ordered sequences, as opposed to properties that only apply to finite sequences. And things can get even worse if we take into account the fact that the difference in the number of marbles in A and B becomes null exactly at $t_b$, i.e. at the first instant after all marbles have already been added. For we then have:

$$\forall t \in (t_1, t_b) : \exists t_v \in \{t_i\} : t_v \leq t < t_{v+1}.$$

Consequently, at instant t the difference d(t) in the number of marbles inside A and B is

$$d(t) = d_v = H^{1+\mathrm{int}(v/H)} + (v \bmod H) - \mathrm{int}(v/H)$$

which obviously increases with v and then with t within $(t_1, t_b)$. How can it finally be $d(t_b) = 0$? Is it unreasonable to suspect that there is something amiss here?

In order to introduce our second variant of Ross's supertask $\{R_i, t_i\}$ (which hardly differs from the original version), consider an ω-ordered collection of identical marbles $\{m_i\}$ labelled with the successive natural numbers, and assume that at each instant $t_i$ of $\{t_i\}$ we add a group of H marbles, labelled from $(i - 1)$ $H + 1$ to iH, to an initially empty box A. In addition, the box A is provided with a mechanism M that removes the marble with the least index from the box while a new set of H marbles is added to the box, including the first set. The mechanism M is set in such a way that it only works within the interval $[t_a, t_b)$. Under such conditions, each marble $m_i$ will be removed at instant $t_i$, an instant at which the box will contain exactly $i(H - 1)$ marbles. Therefore, as the supertask $\{R_i, t_i\}$ progresses, the number of marbles in A varies according to the following strictly increasing ω-ordered sequence of natural numbers

$$\{i(H-1)\} = 1(H-1), 2(H-1), 3(H-1), 4(H-1), \ldots$$

On the one hand, each marble $m_i$ being removed from the box at $t_i$, the one-to-one correspondence $f(t_i) = m_i$ proves that at $t_b$, once the supertask $\{R_i, t_i\}$ is completed, all marbles have been removed from the box. This isn't open to debate. The conclusion to the effect that, at $t_b$, the box A is empty is a direct consequence of a bijection.

On the other hand, let t be any instant within $(t_1, t_b)$, $t_b$ being once again the limit of the sequence $\{t_i\}$. We shall have:

$$\forall t \in (t_1, t_b) : \exists t_v \in \{t_i\} : t_v \leq t < t_{v+1}.$$

Consequently, at instant t, the number n(t) of marbles within the box A will be

$$n(t) = n(t_v) = v(H - 1)$$

which strictly increases with v, and then with t within $(t_1, t_b)$. It is therefore impossible for box A to be empty at any instant within the interval $(t_1, t_b)$. Therefore, taking into account that at $t_b$ no marble has been removed from the box (because at $t_b$ the supertask $\{R_i, t_i\}$ is already completed and the mechanism M is off), the box cannot be empty at $t_b$. This isn't open to debate either. The conclusion to be drawn for this variant of Ross's paradox can only be that box A is and is not empty at $t_b$.

To conclude this section on marbles and boxes, consider the ω-ordered collection of marbles $\{m_i\}$ of supertask $\{R_i, t_i\}$, with the same labelling as before. Let us replace box A by a hollow cylinder C of infinite length with a diameter equal to that of the marbles. Assume that at each instant $t_i$ of $\{t_i\}$ the marble $m_i$ is introduced into the cylinder through its left end. At $t_b$ all the marbles will have been introduced into C. If we now introduce a rigid rod through the left end of C, the rod may hit a marble $m_v$, proving that only a finite number v of marbles have been introduced into C. But it may also be the case that the rod travels the whole length of the cylinder without hitting any marble, as there is no last marble to be hit in the ω-ordered sequence of marbles $\{m_i\}$, but this contradicts the fact that infinitely many marbles were introduced inside the cylinder.

## 11.7 Synchronizing a Supertask

To illustrate the need for H0 in supertasks discussions, a mathematical supertask will now be carried out synchronically with a classical supertask. The classical supertask will be carried out with the collaboration of the infinitely patient guests of Hilbert's Hotel. Let us recall some of the extraordinary properties of this illustrious hotel. Its director has just discovered a new infinitist way of getting rich: he or she demands one euro from $G_1$ (the guest of room $R_1$); $G_1$ recovers his or her euro by demanding one euro from $G_2$ (the guest of room $R_2$); $G_2$ recovers his or her euro by demanding one euro from $G_3$ (the guest of room $R_3$), and so on and so forth. Finally, each guest recovers his or her euro since there is no "last guest" losing his or her money. The deceitful director then demands a second euro from $G_1$ who recovers it again by demanding one euro from $G_2$, who recovers it by demanding one euro from $G_3$, etc. Thousands of euros thus pass from (infinitist) nothingness to the pocket of the fortunate director!

Eccentricities aside, let us assume that the rooms of the hotel are arranged in a unique row divided into two adjacent parts, the left-hand side and the right-hand side. The right-hand side is an ω-ordered sequence of contiguous rooms labelled $R_1$, $R_2$, $R_3$, … from left to right. The left-hand side is also an ω-ordered sequence of

contiguous rooms, labelled … $L_3$, $L_2$, $L_1$ from right to left, and in such a way that $L_1$ is contiguous with $R_1$. Symbolically:

$$HH = \ldots L_5\, L_4\, L_3\, L_2\, L_1\, R_1\, R_2\, R_3\, R_4\, R_5 \ldots$$

In addition to its entrance front door, each HH room has two lateral doors: a left door that communicates with the contiguous room to the left, and a right door that communicates with the contiguous room to the right. We shall also assume that along the interval $[t_a, t_b]$ all lateral doors are open and that all entrance doors of every single room are blocked, so that no guest can leave the hotel. To say that a room $L_i$ or $R_i$ is empty we shall write "$L_i^o$, $R_i^o$," and to say that the guest $G_n$ occupies them we shall write "$L_i^{Gn}$, $R_i^{Gn}$." We shall assume that, initially, each right room $R_i$ is occupied by guest $G_i$, all left rooms being initially empty. So the initial state of HH at $t_a$ will be:

$$HH(t_a) = \ldots L_5^o\ L_4^o\ L_3^o\ L_2^o\ L_1^o\ R_1^{G1}\ R_2^2\ R_3^{G3}\ R_4^{G4}\ R_5^{G5} \ldots$$

Let us now consider the following HH-change: through the left door of his or her room, guest $G_1$ moves to the left empty room contiguous to his or her current room (provided that such an empty room exists) and each guest $G_{i,\, i>1}$ moves through the left door of its current room to the room previously occupied by $G_{i-1}$:

$$HH(t_1) = \ldots L_5^o\ L_4^o\ L_3^o\ L_2^o\ L_1^{G1}\ R_1^{G2}\ R_2^{G3}\ R_3^{G4}\ R_4^{G5}\ R_5^{G6} \ldots$$
$$HH(t_2) = \ldots L_5^o\ L_4^o\ L_3^o\ L_2^{G1}\ L_1^{G2}\ R_1^{G3}\ R_2^{G4}\ R_3^{G5}\ R_4^{G6}\ R_5^{G7} \ldots$$
$$HH(t_3) = \ldots L_5^o\ L_4^o\ L_3^{G1}\ L_2^{G2}\ L_1^{G3}\ R_1^{G4}\ R_2^{G5}\ R_3^{G6}\ R_4^{G7}\ R_5^{G8} \ldots$$

By either induction or Modus Tollens, it can be easily proved that for every natural number v it is possible to carry out the first v HH-change (Theorem 1).

Let now $A_0 = \{a_1, a_2, a, \ldots\}$ be an ω-ordered set, and consider the following ω-ordered sequence of recursive definitions $\{D_i(A_i)\}$ of the sequence of nested sets $\{A_i\}$:

$$i = 1, 2, 3, \ldots D_i(A_i) : A_i = A_{i-1} - \{a_i\}.$$

Let us assume that at each instant $t_i$ of the sequence of instants $\{t_i\}$ the ith definition $D_i$ of $\{D_i(A_i)\}$ is provided and that, at the very same instant $t_i$, the ith HH-change occurs (provided that it can indeed occur). At $t_b$, once the infinitely many successive definitions $D_i$ have been provided (supertask $\{D_i, t_i\}$) and thanks to H0, we shall have a new sequence of nested sets $\{A_i\}$ exhaustively defined as a complete totality. We shall be able to resort to it whenever needed, for instance to prove new theorems. This is typical of standard infinitist mathematics (were it not for the fact that standard infinitist mathematics is not interested in timetabling the steps of ω-ordered sequences of steps).

Things are quite different with $\{HH_i, t_i\}$. As a matter of fact, at $t_b$, and once all possible HH-changes have taken place, all guests mysteriously disappeared from the hotel: $G_n$ being any guest, he or she cannot be in any right room $R_k$ (for any natural number k) nor in any left room $L_p$ (for any natural number p). In the first case, only the first $n - k$ HH-changes would have occurred, while in the second that number would be $p + k - 1$.

It is clear that both results contradict Theorem 1. So, if H0 applies to supertask $\{HH_i, t_i\}$ in the same way that it applies to the recursive definition $\{D_i(A_i)\}$, we have a serious conflict with the ω-ordering derived from the axiom of infinity. In any event, anyone denying that there is a conflict should provide a cogent argument to that effect.

## 11.8 Mathematical Supertasks

Definitions, recursive definitions, procedures and proofs involving infinitely many successive steps are common in contemporary mathematics. In general, mathematicians are not interested in the way these infinitely many steps might be taken. They take it for granted that they are and focus on the results. If these results consist in infinite collections such as sets or sequences, these are construed as completed totalities in agreement with the hypothesis of actual infinity embedded in the axiom of infinity, which in turn implies that infinitely many steps have indeed been taken. Of course, all these infinite definitions, procedures or proofs assume H0. Unless H0 is assumed, once these infinitely many steps of the corresponding definitions, procedures and proofs have been taken, one could be in the odd situation of being unable to assert anything about them: they would be utterly useless. Uninteresting as it may seem from a purely mathematical perspective, we could schedule these infinitely many successive steps in the form of supertasks. These mathematical supertasks have the advantage of being immune to the problem of change. They nevertheless have the same discursive functionality than standard supertasks and could be used, e.g., to examine basic transfinite principles such as ω-order and the axiom of infinity. We shall now introduce some of these mathematical supertasks, although not in detail (for detailed arguments, see Antonio León-Sánchez 2013).

## 11.9 Lost in Exchanges

Let $\{a_i\} = a_1, a_2, a_3, \ldots$ be an ω-ordered sequence and construe it as a table with one row and infinitely many columns both ω-ordered and indexed by the successive natural numbers. Assume now that we successively exchange $a_1$ for the element placed in the next column to the right of $a_1$. Let us call these exchanges "$a_1$-exchanges." After the first n successive $a_1$-exchanges we would have:

$$\mathbf{a_1}, a_2, a_3, a_4, \ldots a_n, a_{n+1} a_{n+2}, a_{n+3} \ldots$$
$$a_2, \mathbf{a_1}, a_3, a_4, \ldots a_n, a_{n+1} a_{n+2}, a_{n+3} \ldots$$
$$a_2, a_3, \mathbf{a_1}, a_4, \ldots a_n, a_{n+1} a_{n+2}, a_{n+3} \ldots$$
$$\ldots$$
$$a_2, a_3, a_4, a_5, \ldots a_n, a_{n+1} \mathbf{a_1}, a_{n+2}, a_{n+3} \ldots$$

It is easy to prove either by induction or Modus Tollens that for any natural number v greater than 0 it is possible to carry out the first v successive $a_1$-exchanges (Theorem 2). Consider now the following supertask $\{a_i, t_i\}$: at each successive instant $t_i$ of $\{t_i\}$, exchange $a_1$ for the element in the next adjacent column to the right of $a_1$ (provided that such a column exists), otherwise stop the supertask. In any event, the supertask will have been achieved at $t_b$. Let v be any natural number and assume that the element $a_1$ is in the vth column of $\{a_i\}$ at $t_b$. If that were the case, the first v $a_1$-exchanges would not have been carried out, which contradicts Theorem 2. Therefore, v being any natural number, we must conclude that the element $a_1$ is no longer an element of $\{a_i\}$ at $t_b$. At $t_b$, $a_1$ has disappeared from the table in spite of the fact that no $a_1$—exchange made it disappear.

## 11.10 The Next Rational

The set $Q^+$ of positive rational numbers in their natural order of precedence is densely ordered: between any two rationals, infinitely many other rationals exist. But $Q^+$ is also denumerable, so it can be put into a one-to-one correspondence f with the set N of natural numbers. This correspondence f induces an ω-order in $Q^+$: $\{q_1, q_2, q_3, \ldots\}$, where $q_n$ is the rational number f(n). Therefore, the set of positive rational numbers may be both densely ordered (between any two rationals infinitely many other rationals exist) and ω-ordered (between any two successive rationals no other rational exists).

This kind of numerical disorder allows us to develop the following argument.

Let x be a rational variable whose initial value is 1, and consider the following sequence $\{D_i(x)\}$ of x redefinitions:

$$i = 1, 2, 3, \ldots :$$
$$\text{If } |q_{i+1} - q_1| < x, \text{ then } D_i(x) : x = |q_{i+1} - q_1|,$$
$$\text{otherwise } D_i(x) : x \text{ remains the same}$$

which redefines "x (for each i = 1, 2, 3, …)" as "$|q_{i+1} - q_1|$ if $|q_{i+1} - q_1|$ is less than the current value of x, $|q_{i+1} - q_1|$ being the absolute value of $q_{i+1} - q_1$, and <being the natural order in $Q^+$."

It is worth noting that all the successive definitions $D_i(x)$ redefine the very same object, i.e. the rational variable x. By contrast, each recursive definition $D_i(A_i)$ of

the above sequence of definitions $\{D_i(A_i)\}$ defines a different object, i.e. the set $A_i$ of the sequence of nested sets $\{A_i\}$. This divergence is crucial: in one case we have an ω-ordered sequence of definitions without a last definition that originates an ω-ordered sequence of nested sets without a last set. In the other case, we also have an ω-ordered sequence of definitions without a last definition, but all these successive definitions define the very same object x, which forces $\{D_i(x)\}$ to leave a permanent trace in the form of the rationality of x: once all possible redefinitions $D_i(x)$ have been provided, the variable x will still be a rational variable, albeit one that has been redefined a certain number of times. Otherwise, H0 would have been violated, and the very same violation could happen to any sequence of successive definitions of the same object or of different objects. The only alternative to this would be to forbid to redefine the same object infinitely many successive times. The prohibition would then have to be added as a new restrictive axiom of transfinite mathematics. For the time being, we shall proceed with our argument about $D_i(x)$ since we are, as a matter of fact, allowed to redefine an object any finite or infinite number of successive times.

By either induction or Modus Tollens, it can easily be proved that for any natural number v, the first v redefinitions of the sequence $\{D_i(x)\}$ can be carried out (Theorem 3). Consider the following supertask $\{D_i(x), t_i\}$: perform $D_i(x)$ at instant $t_i$ if it is possible to perform $D_i(x)$; if not, stop the supertask. In any event, all possible redefinitions $D_i(x)$ will have been carried out at $t_b$. Whatever the value of x at $t_b$ may be, it will be a rational number since x is a rational variable, i.e. one that can only take rational values even though it has been redefined a given number of times. Otherwise, we would be violating H0. We now prove the following two contradictory results with respect to the value of x at $t_b$.

R1: At $t_b$, the rational $q_1 + x$ is not the least rational greater than $q_1$.

*Proof* Q Being densely ordered, the rational number $q_1 + 0.1x$, for instance, is greater than $q_1$ and less than $q_1 + x$. So $q_1 + x$ is not the least rational greater than $q_1$.

R2: At $t_b$, the rational $q_1 + x$ is the least rational greater than $q_1$.

*Proof* Assume it is not. $Q^+$ being ω-ordered, there will be a $q_v$ in $\{q_1, q_2, q_3, \ldots\}$ such that:

$$q_1 < q_v < q_1 + x$$

and then

$$0 < q_v - q_1 < x$$

which implies that the vth redefinition $D_v(x)$ (the one that would have defined x as $q_v - q_1$) has not been provided, a conclusion which contradicts Theorem 3. So $q_1 + x$ is indeed the least rational greater than $q_1$.

## 11.11 Cantor's (1874) Argument

In 1874, Cantor proved that the set A of algebraic numbers and the set Q of rational numbers are both denumerable (Cantor 1874). He also proved in the same paper that the set R of real numbers is not denumerable. This argument is as conclusive as his diagonal proof, although it is far less well known. Cantor's 1874 argument leads to three exhaustive and mutually exclusive possibilities, each of them proving that R is not denumerable. Two out of the three could also be applied to the set Q of rational numbers. So we need to prove that Cantor's 1874 argument always leads to the third option when it is applied to the set of rational numbers, otherwise Q would also be non-denumerable. Unless we prove that this is indeed the case (and the proof promises to be far from obvious), set theory is facing a contradiction with respect to the cardinality of the set of rational numbers. It is hard to believe that neither Cantor nor his infinitist successors ever realized that such a proof is necessary.

We shall resort to a variant of Cantor's 1874 argument to define a new mathematical supertask with conflicting consequences. As noted above, Q can be put into a one-to-one correspondence f with N. So we can define an ω-ordered sequence of rational numbers $\{q_i\} = \{f(i)\}$ that contains all rational numbers. Let (a, b) be any rational interval, and let x, a rational variable whose domain is (a, b) and whose initial value is c, be any element of (a, b). Consider then the following ω-ordered sequence $\{D_i(x)\}$ of successive x redefinitions:

$$
\begin{aligned}
&i = 1, 2, 3, \ldots : \\
&\text{If } q_i \in (a, b) \text{ and } q_i < x, \text{ then } D_i(x) : x = q_i, \\
&\text{otherwise}, D_i(x) : x \text{ remains the same}
\end{aligned}
$$

which compares x with the successive $q_i$ of $\{q_i\}$ within (a, b), and redefines x as $q_i$ each time $q_i$ is in (a, b) and is less than the current value of x. By either induction or Modus Tollens, we may prove that for each natural number v it is possible to provide the first v definitions of the sequence $\{D_i(x)\}$ (Theorem 4).

Assume that, if carrying out $D_i\{x\}$ is possible, it is carried out at instant $t_i$ of the sequence $\{t_i\}$, otherwise stop the supertask $\{D_i(x), t_i\}$. In any case, at $t_b$, all possible redefinitions of the sequence $\{D_i(x)\}$ will have been carried out. According to H0, x will be defined at $t_b$ as a rational number within the interval (a, b), since x is a rational variable whose domain is (a, b) and has been redefined a given number of times. Consider then the rational interval (a, x), and let q be any of its elements. Obviously, q is in (a, b) because (a, x) is a subinterval of (a, b). Yet q cannot be an element of $\{q_i\}$. Let us suppose that it is. In that case, $q = q_v$, for a certain $q_v$ in $\{q_i\}$, and then $q_v < x$ because $q_v$ is in (a, x). But this implies that the vth redefinition $D_v(x)$ has not been carried out and this conclusion contradicts Theorem 4 (notice that $D_v\{x\}$ would have redefined x as $q_v$). We must conclude that the sequence $\{q_i\}$ that contains all rational numbers does not contain all rational numbers.

## 11.12 Cantor's Diagonal Argument

Cantor's diagonal argument is one of the most celebrated and productive arguments in the recent history of logic and mathematics (Cantor 1890–1891). It is a simple and elegant Modus Tollens proving that the set of real numbers is non-denumerable. Despite many criticisms, it is correct. It is relatively common in infinitist discussions to reject an argument because the conclusion of another independent argument contradicts its conclusion. This has been the case with Cantor's diagonal argument, and it is clearly inadmissible because two independent arguments that happen to contradict each other amount to a proof of a contradiction. An argument may be dismissed only when one is able to show where and why *that particular argument* fails.

Cantor's diagonal argument also poses a problem that has not been adequately addressed so far: could the indexed rows of Cantor's table be permuted in such a way that the resulting table defines a rational diagonal (and then a rational antidiagonal)? Clearly, and for the same reasons as in Cantor's 1874 argument, if that were the case, we would be facing a contradiction with respect to the cardinality of the set of rational numbers. As with the alternatives of Cantor's 1874 argument, we need to prove that such a reordering of the rows of Cantor's table is not possible if we want to reject the contradiction (once again, the proof is far from obvious). It is striking that little attention has been paid to this problem.

Our last mathematical supertask is related to Cantor's diagonal argument (in this case, some auxiliary work is nevertheless necessary). To begin with, we need to prove the following theorem of the nth decimal:

For every natural number n there are infinitely many different rationals in (0, 1) with the same decimal $d_n$ in the same nth position of its decimal expansion.

Without going into all the details, the sequence of rationals numbers:

$$
\begin{aligned}
q_1 &= 0.d_1d_2\ldots d_n1\\
q_2 &= 0.d_1d_2\ldots d_n11\\
q_3 &= 0.d_1d_2\ldots d_n111\\
q_4 &= 0.d_1d_2\ldots d_n1111\\
&\ldots
\end{aligned}
$$

and the bijection $f(n) = q_n$ will suffice to prove the theorem.

Let us recall that Cantor's hypothetical indexed table $\{r_i\}$ contains all the real numbers (rational and irrational) within the interval (0, 1), one number in each row $r_i$. Although it isn't necessary as far as the next argument is concerned, Cantor's table could easily be redefined so as to make sure that it contains at least all the rational numbers within (0, 1).

The decimal expansion of rational numbers with a finite decimal expansion will be completed with infinitely many 0s to the right of their last decimal. So, in place

of, say, 0.25, we shall write 0.25000… etc. Finally, we shall say that a row of Cantor's table is n-modular if the nth decimal of its decimal expansion is (n mod 10). For instance:

row $r_1$: 0.60**3**0**5**111022**3**39 … is 3-modular, 5-modular, 13 modular, etc.
row $r_2$: 0.0**2**000**67**1010000 … is 2-modular, 6-modular, 7 modular, etc.
row $r_3$: 0.**11**3000000000000 … is 1-modular, 3-modular, 10 modular, etc.

If the nth row $r_n$ of Cantor's table is n-modular, we shall say that it is D-modular (in the above examples, the rows $r_2$ and $r_3$ are D-modular). If a row $r_i$ is not D-modular, it can be exchanged for any of the following i-modular rows $r_{j,\ j>I}$, provided that such a row exists. Once exchanged, $r_i$ will contain the number in $r_j$ (and vice versa) and it will become D-modular. We call these exchanges "D-exchanges." This is all we need to define the next Cantorian diagonal supertask.

Assume that, at each instant $t_i$ of $\{t_i\}$, the row $r_i$ of Cantor's table is such that:

If $r_i$ is D-modular, it remains unchanged.
If $r_i$ is not D-modular and can be D-exchanged for any following i-modular row $r_{j,\ j>i}$, it is D-exchanged.
If $r_i$ is not D-modular and cannot be D-exchanged, it remains unchanged.

Notice that once a non-D-modular row $r_i$ has been D-exchanged, it becomes D-modular and will remain D-modular (and unaffected by the subsequent D-exchanges) due to the condition $j > i$ (in $r_{j,\ j>i}$) on D-exchanges. At $t_b$, all rows will have been taken into consideration and the supertask will have been completed. Let this supertask be denoted by "$\{r_i, t_i\}$."

We may now prove that, at $t_b$, once $\{r_i, t_i\}$ has been carried out, all rows of Cantor's table are D-modular. Let us assume that they are not, i.e. let us assume that there is a row $r_n$ at $t_b$ that is not D-modular. As a consequence of ω-asymmetry, $r_n$ has a finite number $(n - 1)$ of preceding rows and an infinite number of succeeding rows. This implies that $r_n$ can only be preceded by a finite number of n-modular rows. According to the theorem of the nth decimal, there are infinitely many rationals with the same decimal (n mod 10) in the same nth position, since all n-modular rows have the same decimal (n mod 10) in the same nth position of their decimal expansion. In other words, there are infinitely many n-modular rows, of which only a finite number precede $r_n$. Consequently, $r_n$ is succeeded by infinitely many n-modular rows and hence had to be D-exchanged for any one of them. Therefore $r_n$ must be D-modular. We must conclude that, once the supertask $\{r_i, t_i\}$ has been carried out, all rows of Cantor's table are D-modular. As in the case of Cantor's diagonal argument, this one is also a simple Modus Tollens. The reader can easily prove that supertask $\{r_i, t_i\}$ leads to other conflicting results such as the disappearance of infinitely many rows from the table.

If all the rows of Cantor's table become D-modular, the new diagonal of the table will be a periodic rational number within (0, 1), whose period is 1234567890, i.e. the rational number 0.123456789012345678901234567890…

From this diagonalization, we may define infinitely many rational antidiagonals, e.g. the periodic rationals within (0, 1) of periods 0123456789, 45, 21, etc. For the same reasons as with irrational antidiagonals, each of these rational antidiagonals would prove that the set of rationals within (0, 1) is not denumerable. Therefore, we would have a fundamental contradiction in set theory: the set of rational numbers would be both denumerable and non denumerable. The alternative to this contradictory conclusion would be a violation of H0, so that, for example, at $t_b$, either the rows of Cantor's table are no longer real numbers within (0, 1), or they are arbitrarily permuted so that we cannot be sure that all of its rows are D-modular, etc. Obviously, the same final arbitrary effects could be expected with any definition, procedure or proof involving infinitely many successive steps, and transfinite mathematics would then be a meaningless affair.

## 11.13 Partitions à la Cantor

The following summarized argument is not a supertask but a mathematical procedure with infinitely many steps, expressed in the compact form of computer language. It illustrates another way of posing problems related to the hypothesis of actual infinity. It is inspired by Cantor's Ternary Set (Cantor's power) and by Cantor's argument about the partition of the real line in Cantor (1885).

Let A be the real interval (0, 1) and X a set of indexes whose elements will be referred to as a, b, c, d, … and whose cardinality is $2^{\aleph_0}$. Let u and v be two real variables and f a one-to-one correspondence between A and X.

Consider the following procedure P:

u = 0 Used to define the left end of the successive intervals.
v = 0 Used to define the right end of the successive intervals.

When $A \neq \varnothing$:

X = X − {a} a (b, c, d, …) is any element of X.
A = A − {f(a)} Remove f(a) from A.
v = v + f(a) Right end of the interval.
If v is in R then R is the set of proper real numbers.
$(x_a, y_a] = (u, v]$ Define a new real interval.

Otherwise:

Exit Loop Two real numbers whose sum is not a proper real number.

End if

u = v Left end of the next adjacent and disjoint interval.

Loop.

Since the sum of any two real numbers is a real number, P exhausts the sets I and A and defines a partition $T = \{(x_a, y_a], (x_b, y_b], (x_c, y_c], \ldots\}$ on the real line, each of whose intervals $(x_h, y_h]$ defines a different real number $r_h = y_h - x_h$. We may then prove that $g((x_h, y_h]) = r_h$ is a one-to-one correspondence between T and (0, 1). Thus, by selecting a rational number $q_h$ within each $(x_h, y_h]$, we would have a non-denumerable sequence of different rational numbers. As with Cantor's 1874 argument and Cantor's diagonal argument, this new conclusion also points to a possible contradiction with respect to the cardinality of the set of rational numbers.

## 11.14 Infinity and Physics

While many authors believe in the formal consistency of supertasks, a (much smaller) number of them believe in the physical possibility of actually carrying out supertasks. In this latter group, there are those who believe that supertasks could be carried out in infinite intervals of time that are perceived as finite intervals thanks to the relativistic dilation of time (these are the so-called "bifurcated supertasks"). As we shall see now, ω-asymmetry and quantum mechanics add new difficulties to the possibility of actually carrying out (or even observing) a supertask in a finite interval of time.

Let $t_p$ be Planck time ($5.39 \times 10^{-44}$ s) and consider the time interval $(t_b - t_p, t_b)$. Due to ω-asymmetry, at any instant t within $(t_b - t_p, t_b)$, only a finite number of actions have been carried out and there still remains an infinite number of actions to be completed. Since $t_b$ is the limit of $\{t_i\}$, there will exist an instant $t_v$ in $\{t_i\}$ such that $t_v \leq t < t_{v+1}$ and hence, at instant t, only a finite number v of actions will have been carried out. So infinitely many actions would have to be completed within an interval of time far smaller than Planck time.

A simple exercise in differential calculus proves that, assuming Heisenberg principle of uncertainty, Planck length and Planck time are, respectively, the shortest length and least time to be measured in physical terms. So we could never verify in these terms that a supertask has been carried out in a finite interval of time simply because infinitely many actions would have to be carried out in less than Planck time. Furthermore, as most physicists now suspect, physical laws might no longer hold beyond Planck scale. It is also quite plausible that nothing in the physical world can last a time shorter than Planck time. This adds some additional difficulties to the claim that supertasks could be physically carried out.

We have only been concerned here with conceptual ω-supertasks. Our purpose wasn't to enlist physics in supertask theory but to illustrate the way in which supertasks could be used to call into question the axiom of infinity. This questioning is nevertheless of great interest for the experimental sciences, e.g. physics.

Apart from ω-asymmetry, ω-inconsistency also applies to physical supertasks: as far as the number of actions to be carried out is concerned, that number can only take two values: $\aleph_o$ and 0. The only solution to this dichotomy is that infinitely

many successive actions be carried out simultaneously. But successive actions cannot be carried out simultaneously in physical terms.

The mathematical infinite, on the other hand, is anything but a trivial matter. Consider its impact on physics. The points of the infinitist continuum of the real numbers, for instance, are of capital importance in physics: think of point masses, point particles, point charges, etc. The special theory of relativity, one of the more successful theories of modern physics, is a physical theory of the spacetime continuum. Theoretical physics relies almost exclusively on infinitist mathematics. But, as we shall see, the continuum may well not be the best model of the physical world, particularly when we approach ultramicroscopic scales such as the Planck scale. The persistent problems that physicists have been concerned with for more than fifty years suggest that the physical world could be discrete and discontinuous, i.e., digital.

For many contemporary physicists, the persistent incompatibility between quantum mechanics and general relativity is a consequence of the lack of discreteness of the continuum-based models (see Majid 2008). They suspect that space and time are not continuous but discrete (i.e. composed of indivisible minima) and that the granular fabric of spacetime could be the meeting place for the two fundamental physical theories. An increasing number of theoretical and experimental research is now trying to prove the discrete nature of space and time. As we shall also see, the search for violations of Lorentz symmetry at Planck scale (the plausible granular scale of the physical world) is now becoming an active area of theoretical and experimental research (see Smolin 2007; Maudlin 2011). The continuum being a formal descendent of the axiom of infinity, we must insist on the importance of reexamining the formal consistency of the actual infinity hypothesis, if only for the fact that, if it were inconsistent, so would all continuums. Most physicists have not considered this possibility.

Some of our conclusions on supertasks and mathematical supertasks suggest that the hypothesis of actual infinity embedded in the axiom of infinity may be inconsistent. If that were the case, more than a century of mathematics would have to be revised. As a consequence, a new type of discrete (digital) mathematics would have to be developed, including discrete analysis and discrete geometry (more on this below). Physics would also be affected by these conclusions, albeit in a different way. This is, at least, what the following points seem to indicate:

1. Although theoretical physics relies on infinitist mathematics, experimental physics deals with digital results. Even when dubbed analog, observations and measurements are always discrete, truncated to a small number of digits.
2. Infinities that emerge in physical equations have to be removed in order to avoid the unsolvable problems they invariably lead to, e.g. in the standard model of particles (renormalization).
3. All physical magnitudes seem to be of a discrete nature, with indivisible minima. Even space and time are also suspected of being composed of indivisible minima.

4. Nothing we have ever observed, measured or divided is infinite. In physics, infinity could just be a "manner of speaking."
5. The suspected digital scale of nature could be the Planck scale. As experimental and theoretical physics get closer to this scale, some problems emerge. Most physicists suspect that it is not possible to find a solution to these problems within our current analog models. (We shall deal with one of them in this section.)

If nature is indeed discrete in all its observables, the analog mathematics of the continuum may very well not be the most appropriate instrument to deal with the physical world. When we examine the world independently of its discrete scale of minima, the analog mathematics of the continuum works quite well. Problems surface when we approach that scale.

Before approaching the hypothetical digital scale of the physical world (and the problems it poses for some well-established physical theories), let us recall the n-expofactorial numbers of unimaginable size such as $9^{!9}$. We certainly may *define* finite numbers that large, but to suppose that they are in any way meaningful is quite another matter. Suppose there is indeed a real number with $9^{!9}$ decimals in its decimal expansion. Suppose furthermore that all the physical constants needed to explain the universe are real numbers with $9^{!9}$ decimals (and $9^{!9}$ is still very small compared with $\aleph_o$). Wouldn't such a universe be utterly absurd? On the other hand, all periodic rational numbers and all irrational numbers have an ω-ordered sequence of $\aleph_o$ decimals in their decimal expansions. Moreover, according to infinitist orthodoxy, these numbers exist as a complete totality. Most of the irrational numbers within (0, 1) are supposed to have an infinite (ω-ordered) and random sequence of decimals. Being infinite and random as they are, it would be easy to prove that each of these sequences contains an encoded version of all known texts written by humans from the Neolithic up to now. What is even more incredible is that every one of them would also contain sequences of the same decimal repeated, e.g., $99^{!99}$ times. These are unavoidable consequences of the alliance of randomness with actual infinity.

As we know, $\aleph_o$, the cardinal of the set of natural numbers, is the smallest number greater than all finite natural numbers. The problem with $\aleph_o$ is that its definition is not related to the operational definition of natural numbers via the successor set. $\aleph_o$ is not the successor of any finite natural number simply because there is no last natural number to be succeeded by $\aleph_o$. Cantor proved that $\aleph_o$ is not a natural number because $\aleph_o = \aleph_o + 1$, while every natural number n satisfies $n \neq n + 1$. He then proved that $\aleph_o$ is greater than all finite cardinals because, for any finite cardinal n, the set {1, 2, … n} is not equivalent to N although it is a proper part of it. He also proved that it is the least cardinal greater than all finite cardinals (see Cantor 1955: Theorems A and B, epigraph 6). So $\aleph_o$ is the limit of all strictly increasing and ω-ordered sequences of natural numbers; but $\aleph_o$ does not play in physics the fundamental role that it plays in set theory.

The next transfinite cardinals are $2^{\aleph_o}$ and $\aleph_1$. The former is the cardinal, among others, of the set of real numbers, i.e. the power of the continuum. The latter is the

cardinal of the set of all ordinals whose sets have the same cardinal $\aleph_o$. We do not know if $2^{\aleph_o} = \aleph_1$ (Continuum Hypothesis). The sequence of powers ($2^{\aleph_o}$, $2^{2^{\aleph_o}}$, $2^{2^{2^{\aleph_o}}}$, …) and the sequence of alephs ($\aleph_1$, $\aleph_2$, $\aleph_3$, …) yield an unending number of infinities of an increasing infinitude that leads towards absolute infinity, which is infinite to the point of being inconsistent, at least, if Cantor is to be believed, for our limited human minds. All of them, except the power of the continuum, are irrelevant to physics.

The spacetime continuum is grounded on the continuum of the real numbers, whose cardinal is $2^{\aleph_o}$ (in physics literature, even under the signature of Nobel laureates, it is not unusual to meet with the erroneous assertion that this cardinal is $\aleph_1$. After all, transfinite arithmetic may not be essential for doing physics). As is well known, the continuum of the real numbers is so infinite that a straight line segment of Planck length ($1.62 \times 10^{-35}$ m) has the same number of points as the whole tridimensional visible universe, which, bijections aside, is rather enigmatic from a purely physical point of view (think, for instance, of virtual quantum particles). Bijections and ellipses thus form an unlikely, not to say dangerous, couple.

The physical theory which turns out to be directly concerned with the hypothesis of actual infinity is the special theory of relativity; it was founded on the contention that space and time are unified into a four-dimensional continuum called spacetime, this continuum being the infinitist continuum of the real numbers. As is well known, Einstein's theory refines Newtonian mechanics for velocities approaching the speed of light. The special theory of relativity has been satisfactorily confirmed by experiments and observations. The question is: is there any scale of nature at which the theory will also need to be refined, or is it, as such, the ultimate theory? The question matters here because space and time could be discrete rather than continuous, and if they are, some aspects of the special theory of relativity would have to be modified.

Since the beginning of the 21st century there is a growing interest in the search for violations of Lorentz symmetry at the Planck scale. Although this scale was intended to define a metric reference independently of arbitrary definitions of unities for mass, length and time, the interest in Planck's scale has gone far beyond its original metric objectives. It is considered an appropriate candidate for a definition of the granularity (discreteness) of space and time. The Planck scale is defined by a set of universal constants: Planck mass $m_p$, Planck length $l_p$ and Planck time $t_p$ (Planck energy, Planck charge and Planck temperature may also be included). Planck mass, Planck length and Planck time are defined in terms of three universal constants: h (Planck constant), c (the speed of light in the vacuum) and G (constant of gravitation). The universality of the Planck constants poses some significant problems for Lorentz transformation.

The first principle of the special theory of relativity asserts that the laws of physics are universal, i.e. that they are the same in all inertial reference frames. The universality of physical laws implies the universality of the physical constants involved in their mathematical formulations. In addition, if A and B are two universal constants, their algebraic combinations must also be universally constant. For instance, $m_o$ (the magnetic permeability of the vacuum) and $e_o$ (the electric

permittivity of the vacuum) being two universal constants, their algebraic combination $(m_o e_o)^{-1/2}$ is also a universal constant (in this case, the speed c of light in the vacuum). For the same reason, $l_p$, $t_p$ and $m_p$, which are also defined as algebraic combinations of three universal constants (h, c and G in all three cases), can only be universal constants in all reference frames. If that were not the case, say because a given algebraic combination f(h, c, G) had changed with relative motion, the change could only be due to a change in at least one of these three universal constants (provided that real numbers and algebraic operations do not change with relative motion). Thus, at least one out of the three universal constants would change with relative motion: it would not be the universal constant it was assumed to be.

The problem is that Lorentz transformation does not preserve the universality of certain algebraic combinations of h, c, and G. In particular, it does not preserve the universality of Planck length, Planck time and Planck mass. As Smolin pointed out (Smolin 2007), it is astonishing that the problem of the relativity of the universal constants hasn't been posed until the very beginning of the 21st century. Amelino-Camelia proposed a solution that is now known as Doubly Special Relativity, or Deformed Special Relativity (DSR for short) (see Amelino-Camelia 2001). In addition to the speed of light as universal constant, DSR includes two additional universal constants (independent of relative motion): a maximum energy (Planck energy) and a minimum length (Planck length). The theory has now several variants, e.g. DSR II (see Magueijo; Smolin 2003). Not surprisingly, DSRs have not been enthusiastically welcomed.

DSR and its successive variants have built upon the same infinitist mathematics of the continuum as other physical theories. The problem here is that, at Planck scale, we plausibly approach the discrete scale of nature, where the continuum-based mathematics may no longer be the appropriate instrument. The relevance of the formal consistency of the hypothesis of actual infinity becomes striking precisely at this point: if that hypothesis turned out to be inconsistent, so would all continuums formally derived from it, and we would be forced to develop a new discrete mathematics more attuned to the physical world (the branch of mathematics we usually call "discrete mathematics" has nothing to do with this issue). Besides, and for the first time in the history of logic and mathematics, we have at our disposal two productive instruments to dispute that foundational hypothesis: ω, the least transfinite ordinal, with its wealth of asymmetries, dichotomies and possible inconsistencies, and supertask theory, which provides an appropriate scenario by means of which arguments may be represented.

## 11.15 Platonism and Biology

From a physical point of view, an object exists just in case it can interact with other objects in such a way that their states are modified as a consequence of the interaction. It is through interactions that we can detect the existence of physical, i.e.

spacio-temporal, objects. The different branches of physics and other experimental sciences study different types of interactions, from the simple change in the trajectory of a photon to a chemical reaction or a galactic collision. We usually call these "dynamic interactions" because energy is always involved. As far as we know, they are always governed by the same set of universal laws.

Living beings introduce another type of interaction into the physical world, so-called "infodynamic interactions," in which arbitrary signals and codes are involved. Infodynamic interactions modify the state of the receiving objects in such a way that it is not always possible to detect the changes from the physical laws proper, but only from the complex evolutionary and reproductive history of each organism. Obviously, all objects involved in infodynamic interactions are physical objects subject to dynamic interactions with the rest of the world. Apart from being arbitrary, infodynamic interactions are also teleonomic, the objective in most instances being either directly or indirectly related to reproduction (which also includes survival), the universal goal of all living beings.

Living beings survive and reproduce in a physical environment governed by a set of universal physical laws. It isn't surprising that living beings behave in harmony with such laws in order to survive and reproduce, and that nature's regularity has eventually been captured in genetic, epigenetic and neurological terms. In other words, we shouldn't be surprised that we have acknowledged the fundamental laws of logic and proved able to develop formal systems which provide successful models of the physical world.

This natural harmony between our capacities for abstraction and the formal coherence of the physical world have led some to idealize the ontological status of mathematical objects and defend the philosophical doctrine of platonism. But let us recall that mathematical objects have also played an essential role in many *erroneous* physical theories. Moreover, these objects play no role whatsoever in our account of most chemical, geological and biological phenomena, let alone psychological, or sociological ones. Finally, if the above arguments on the hypothesis of actual infinity turn out to be conclusive, and the hypothesis is therefore proven inconsistent, platonism could no longer be defendable: it would no longer make sense because the sequence of natural numbers could only be *potentially* infinite. In other words, natural numbers, the simplest mathematical objects, could only be the result of successive *mental* recursive operations.

Platonism claims that mathematical objects exist in some more profound sense than physical objects. Notwithstanding, the only method available to us to test whether or not some object exists resorts of its (purported) dynamic or infodynamic interactions with other objects. Since non spatio-temporal objects (e.g. abstract objects such as numbers) may not be subjected to this objective physical test, no causal relation can be established with them, which inevitably brings us back to Benacerraf's epistemological argument against platonism. To overcome this difficulty, the platonist is ready to make another claim, i.e. the epistemological claim to the effect that we can have access to them by means of a cognitive ability dubbed intellectual or mathematical "intuition." Let us investigate this contention from the perspective of neuroscience.

To paraphrase Zeki, the organization and laws of the brain dictate all human activity: there can be no genuine epistemology of mathematics unless it is neurobiologically based (Zeki 1995). The kind of mathematical intuition that platonism resorts to in order to argue in favor of a contact between spatio-temporal and non-spatio-temporal objects is incompatible with neuronal processing of information.

We tend to think of vision, hearing and perceptual experience in general as if our brain worked like a camera, capturing external reality, as it were, in a shot. But what the brain does is much more complex and counter-intuitive: it reacts to stimuli, and selects and processes them in different parts of the brain in order to integrate the information it has gathered and to give rise to a unitary conscious experience.

The brain, in a nutshell *constructs* perceptions. It processes sensory outputs, categorizes them, makes generalizations and abstractions that allow us to represent reality. Plato believed that objects are derived from abstraction, but neuroscience has taught us that the derivation takes places the other way around. This capacity for abstraction and concept formation is primitive, allowing us to attain knowledge by means of a construction of all objects of perception.

The importance of such neurological findings cannot be underestimated: there must be enough neurons to represent whatever exists. In other words, if knowing consisted of looking things up in a mental repertoire in which each and every object had been previously registered, the repertoire would be infinite, whereas our physical memory is limited. We may perhaps discuss whether or not there are infinitely many numbers but it may not be doubted that the number of neurons we possess as individuals is finite. Fortunately for us, the process of abstraction saves a lot of energy, neuronally speaking.

When, e.g., we perceive an orange, our preconscious experience is fragmented. The brain processes separately the colour, size, shape, and smell of the object in order for a conscious experience of "perceiving an orange" to emerge at all, i.e. a perception in which all these distinct features are perceived unitarily (unless one suffers from brain injury or merely hallucinates an orange). This is how "an orange becomes all the oranges that exist in the world," and this is how abstraction and generalization take place in the neurophysiological realm (see Mora 2007).

Let us pick up another example to illustrate why anatomic acquisition of knowledge is so efficient. Imagine a set of, say, fifty alphabetical symbols. With such a small number of symbols we could create many languages, in which we could write and tell a vast number of stories. Suppose, on the contrary, that every time we want to tell a story we have to start from scratch and create different alphabets (different sets of letters). This would indeed be very costly in terms of time and energy. Reusing and combining the same letters seems to be a much wiser strategy.

To cope with the problem of limited memory, evolution has selected brains that are able to identify common features shared by distinct objects without having to previously register all of them. Intuition, if there is any such thing, probably results from the intrinsic activity of the brain, i.e. the default mode network discovered by Raichle and his colleagues (Raichle 2006), which is believed to play a major role

in brain function. We know that intuition, whether mathematical or otherwise, is part of the overall machinery by which the brain acquires knowledge. As such, it must be subjected to the same organization and rules that are being investigated by an overwhelming amount of experimental studies.

## References

Amelino-Camelia, G. (2001). Testable scenario for relativity with minimum length. *Physics Letter, B510*, 255–263.

Benacerraf, P. (1962). Tasks, super-tasks, and modern eleatics. *The Journal of Philosophy, 59*(24), 765–784.

Black, M. (1950–1951). Achilles and the tortoise. *Analysis, 11*(5), 91–101.

Cantor, G. F. (1874). Über eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen. *Journal für die reine und angewandte Mathematik, 77*, 258–262.

Cantor, G. F. (1885). Über verschiedene Theoreme aus der Theorie der Punktmengen in einem n-fach ausgedehnten stetigen Raume Gn. *Acta Mathematica, 7*, 105–124.

Cantor, G. F. (1890–1891). Über eine elementare Frage der Mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereiningung*, Band I, 75–78.

Cantor, G. F. (1955). *Contributions to the founding of the theory of transfinite numbers* (Beiträge zur Begründung der transfiniten Mengenlehre, parts I and II, Philip E. B. Jourdain, English Trans.). New York: Dover.

Clark, P., & Read, S. (1984). Hypertasks. *Synthese, 61*(3), 387–390.

Hallet, M. (1984). *Cantorian set theory and limitation of size*. Oxford: Oxford UP.

Antonio León-Sánchez (2013). *Infinity at stake—Selected arguments on the actual infinity hypothesis*. Madrid: Bubok Publishing.

Magueijo, J., & Smolin, L. (2003). Generalized Lorentz invariance with an invariant energy scale. *Physical Review D, 67*(4), 044017.

Majid, S. (2008). Quantum space-time and physical reality. In S. Majid (Ed.), *On space and time* (pp. 56–140). Cambridge: Cambridge UP.

Maudlin, T. (2011). *Quantum non-locality and relativity*. Oxford: Wiley-Blackwell.

Moore, A. W. (2001). *The infinite*. New York: Routledge.

Mora, F. (2007). *Como funciona el cerebro*. Madrid: Alianza.

Pickover, C. A. (1995). *Keys to infinity*. New York: Wiley.

Raichle, M. E. (2006). The brain's dark energy. *Science, 319*, 1249–1250.

Smolin, L. (2007). *The trouble with physics*. London: Penguin Books.

Taylor, R. (1951). Mr Black on temporal paradoxes. *Analysis, 12*(2), 38–44.

Thomson, J. F. (1954–1955). Tasks and super-tasks. *Analysis, 15*(1), 1–13.

Watling, J. (1952). The sum of an infinite series. *Analysis, 11*(2), 39–46.

Zeki, S. (1995). *Una vision del cerebro*. Barcelona: Ariel.

# Part IV
# Retrospection: Mathematical Truth, Mathematical Objects

# Chapter 12
# Mathematical Truth (1968 Version)

**Paul Benacerraf**

## 12.1 Introduction

Contrary to what you might expect from its title, this paper is on the concept of mathematical truth. I will not present an analysis—but I will try to indicate the direction, or directions, I think analyses should follow. I think that many traditional attempts to account for truth, meaning, and knowledge in mathematics are misguided and doomed to failure. Analysis of these failures suggests that the philosopher who wishes to give an account of these notions faces a dilemma. I will describe this dilemma. But my principal aim is to shift the discussion in the philosophy of mathematics *away* from certain traditional lines (which I take to be bankrupt), and onto some new ones. I fear that nothing I say will itself be new for I'm sure everything I say has already been denied in print, probably by Hilary Putnam.

Virtually every question in the philosophy of mathematics depends for its answer on how we view mathematical truth. For example, discussions of the philosophical import of various metamathematical results (Gödel's incompleteness theorems, Cohen's and Gödel's independence proofs) often proceed by asking if arithmetic (set-theoretic) propositions, if undecidable in some standard systems are true (or false) nevertheless. Thus, in the arithmetic case, one side will press the view that Gödel's (Rosser's) discovery that certain axiomatic systems were either incomplete or inconsistent forced one to relativize the notion of arithmetic truth to individual systems—to accept as many different conceptions of arithmetic truth as

Minor revisions of style have been made by Paul Benacerraf in 2014 and 2015. *Editor's note.*

P. Benacerraf (✉)
James S. McDonnell Distinguished University Professor of Philosophy, Emeritus, Princeton, USA
e-mail: paulbena@princeton.edu



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_12

there are different formal systems of arithmetic. Opponents of such a multiplicity of concepts see no unclarity in what is for them "*the*" notion of arithmetic truth. They take Gödel's remarkable achievement as showing that for no formal system of arithmetic does theoremhood coincide with truth. Clearly as many variations on those themes as are possible have been played. Some such divergent views on the import of Gödel's results usually stem directly from different conceptions of arithmetic truth: in the first case, theoremhood in some specified or specifiable formal system is the determinant characteristic, while in the second, it is some principle implying that of every pair consisting of a closed sentence and its negation exactly one member must be true. Quite often, a Tarski-type truth definition in some richer language, usually some set theory, is lurking in the background. More often than not, there is simply the view that every closed sentence is "meaningful" and makes some definite claim which must therefore be true or false. But if we understand negation in the usual way, of every pair consisting of a sentence and its negation, exactly one member must be true and exactly one must be false. The view can be embellished a bit, but not transformed, by making reference to the grammatical forms in question (names, predicates, quantifiers, etc.).

The case of set theory is considerably more complicated; two factors intrude to prevent it from assuming the relative pastoral calm and beauty that attends arithmetic. First, there are the set-theoretical paradoxes discovered around the turn of the 20th century which, despite the current vogue decrying the "Catastrophe Theory" of set-theoretic truth, forced a radical revision in the Fregean concept of set—a revision which detractors of the Catastrophe Theory will claim had been obvious all along—but a radical revision nevertheless.[1] Number theory has not needed to recover from such conceptual illness, no matter how minor you may consider this one to be. The second complication, though perhaps not a defect of birth, is owed to something discovered by Paul Cohen. Cohen showed that the Axiom of Choice and the Continuum Hypothesis are independent from the usually accepted axioms of set theory. This is an *incompleteness* result reaching beyond the one Gödel obtained for arithmetic (which applies to standard formalizations of set theory as well), with what may be more serious conceptual consequences, for the propositions proved independent by Cohen are in a sense more central to the theory. They represent live and debated questions at the very heart of the theory of sets, perhaps genuine branching points in the theory, analogously to the way the independence of the parallels postulate provided a branching point in geometry. The future could see the parallel development of Cantorian and non-Cantorian set theories, according as they accept or reject the Continuum Hypothesis, with neither being given priority. Most practitioners (Gödel and Cohen, for example) seem to feel that these issues will eventually be decided by the adoption of new and evident axioms which characterize further our conception of set—where the emphasis is now upon a *further characterization* rather than an *extension* of our present concept (not a pellucid distinction).

Such is not the case with the class of propositions involved in Gödel's result on arithmetic, for two reasons: (1) for any pair of undecidable Gödel sentences, there is more general agreement concerning *which* of the two should be added to form

extensions of the basic first order arithmetic, and (2) there is a tendency to regard undecidable arithmetic sentences from a *set-theoretic* point of view, and thus to decide them by set-theoretic means—as is evidenced by the following comment of Cohen's:

> Indeed we can postulate as a rather vague article of faith that any statement in arithmetic is decidable in "normal" set theory, i.e. by some recognizable axiom of infinity. This is of course the case with the undecidable statements of Gödel's theorem which are immediately decidable in higher systems.
>
> Cohen (1966: 152)

Whereas the force of the second factor is to relativize arithmetic truth to set-theoretic truth, the independent existence of the first factor provides a minimal testing ground for these additional set-theoretic axioms: they must give the "right" answer to the number-theoretic questions. The difference lies in that whereas there might conceivably exist mathematicians who would investigate "Cantorian" set theories (i.e. ones which imply GCH) and even advocate their adoption, it is unthinkable, as a matter of mathematical anthropology, that anyone seriously contemplate the adoption of systems of arithmetic which take the "wrong" branching at Gödel sentences. Despite these differences, however, the several camps seem once more to form on the basis of what they would take to be the determinants of mathematical (in this case, set-theoretic) truth.

For, one who believes in the possibility of the parallel development of Cantorian and non-Cantorian set theories is likely to do so because she takes the axioms of set theory (e.g. the ZF axioms) to constitute an implicit definition of the notion of set—or of *a* notion of set shown by Cohen to fall short of determining the truth-value of either AC or GCH. I.e., what doesn't follow from these axioms isn't true of sets *thus conceived*—and what is not first order derivable from those axioms doesn't follow from them. Conversely, the "implicit definition" view just mentioned seems to commit its proponent to denying that propositions independent of the axioms have a truth-value at all.

There are many infirmities inherent in such a position, but this is not the place to uncover them in detail. I will simply mention two: (1) It is not usually noticed that excluded middle must be rejected if one is to avoid asserting, for undecidable $S$, $(S \vee \sim S)$, while denying a truth-value to $S$ and to $\sim S$ (otherwise, disjunction must be reinterpreted); (2) It is unclear what concept of logical consequence is being employed when what is at issue is the very theory (set theory) in terms of which the notion of logical consequence is customarily explained. But we will return to this point in another connection. In any event, the branching view seems to have as its rationale a view of set-theoretic truth which identifies a concept of set with the axioms which characterize that concept. Non-equivalent sets of axioms yield different concepts, and something is true under a given conception of set if and only if it is derivable from axioms individuating that conception.[2]

On the other hand, a number of positions seem to be open to those who *deny* that the axioms are implicit definitions of the associated concepts of set. It will suffice

for our immediate purposes to distinguish roughly two: first, there are those who feel that in writing axioms of set theory one is describing some well determined reality—in such a way that each closed formula, whether a theorem or not, has a truth-value determined by how accurately it represents that reality. Secondly, there are those who feel that what is being axiomatized is a (perhaps somewhat vague) underlying conception of set. The axioms we write may or may not correctly characterize sets thus conceived—on *this* score there is no difference with the immediately preceding view—but nor is there any assumption that the conception we are trying to characterize is determinate with regard to every closed sentence that can be formulated in the vocabulary. Holders of this latter view might admit of possible indeterminacies in our conception but need not claim that finding propositions independent of present axioms is *proof* of such indeterminacies: it might at best be proof of our imperfect intuitions concerning the conceptions which activate us. For both of these groups, the prescribed method of research is to explore the consequences of new axioms (whose truth or falsity may not be immediately evident), flooding the axioms with the light of our intuitions based on matters concerning which we feel more secure. The former group thinks of this process as one of discovery, always, while the latter countenances the possibility that propositions might be found which are independent of accepted axioms and with respect to which our intuitions are totally indeterminate, even in the aforementioned "inductive" sense. Such propositions would be without a truth-value (and the problem with excluded middle would arise). However, in either case, what determines the relevant positions on the significance of the independence results is how the axioms were viewed in the first place.[3]

I think these few examples will suffice to point out the pervasiveness and importance of the problem which forms the subject of this paper: how may truth in mathematics be explained? I shall now try to sharpen the question by setting down two conditions which I feel any acceptable answer must meet. It will be the burden of the balance of this paper to describe the general kinds of answer traditionally given to the problem of mathematical truth and argue that each kind satisfies exactly one of the two conditions I set down—in each case apparently at the cost of violating the other. I will thus present a dilemma. There are two conditions which any satisfactory account of mathematical truth must satisfy, but existing accounts each seem to satisfy one of the conditions at the expense of the other. *Must* this be so? Is there an account which meets *both* conditions? Or is there an argument that would suggest that one of the conditions is unreasonable? I conceive of this paper as an invitation to escape the dilemma I pose—not as an attempt to prove that the questions I ask are unanswerable.

Before getting on with the job, I must apologize and beg your indulgence. Much of what I have said and will say will seem extremely vague, perhaps hopelessly so. Still, I think that despite the vagueness some sense, and perhaps even some truth can be eked out.

## 12.2 The Problem

For illustrative purposes, consider the following two sentences:

(1) There are at least three large cities between New York and Chicago.
(2) There are at least three prime numbers between 17 and 43.

Do they have the same logico-grammatical form? Or, more specifically, are they both of the form

(3) There are at least three *FG*'s each of which bears *R* to *a* and *b*,

where "there are at least three" is a normal numerical quantifier, "*F*" and "*G*" are to be replaced by one-place predicates, "*R*" by a three-place predicate, and "*a*" and "*b*" by names of elements of the universe of discourse? What are the truth conditions of (1) and (2)? Barring possible vagueness in the term "large cities" it seems fairly clear that (1) is of the form (3) and thus that it will be true if and only if what are named by the expressions replacing "*a*" and "*b*" ("Chicago" and "New York," respectively) bear the relation designated by the expression replacing "*R*" ("surrounding") to at least three elements of the domain of discourse of the quantifier which satisfy the predicates replacing "*F*" and "*G*" ("large" and "city," respectively).[4] This is what a truth definition *à la* Tarski would tell us. And I think that is right. Thus, if (1) is true it is because certain cities stand in a certain relation to one another, etc. But what of (2)? May we use (3) as a guide to spell out the conditions of its truth? Does (2) contain names where (1) does? Is (2) true if and only if what are named by the expressions replacing "*a*" and "*b*" ("17" and "43," respectively) bear the relation replacing "*R*" ("surrounding") to at least three elements of the domain of discourse of the quantifier which satisfy the predicates replacing "*F*" and "*G*" ("prime" and "number," respectively)? Or is an entirely different kind of answer to be given concerning the conditions of its truth? Certainly the history of the subject (the philosophy of mathematics) has seen other answers given. Some (including one of my past selves),[5] reluctant to face the consequences of such an overtly platonistic account, have shied away from supposing that numerals are names and thus, by implication, that (2) is of the form (3). (I sympathize.) Indeed, on some such accounts, the truth conditions of arithmetic sentences are given as their derivability from specified sets of axioms. When coupled with the desire to have each closed sentence of arithmetic receive a truth-value, such accounts were torpedoed by the incompleteness theorems. They could be rescued at least to *internal* consistency either by liberalizing what counted as derivability (e.g. including an ω-rule in permissible derivations) or by abandoning the desire for completeness. For lack of a better term, I will call such views "combinatorial" views of the determinants of mathematical truth. The leading idea of combinatorial views is that what makes for mathematical truth is certain combinatorial (generally, proof-theoretic) facts about the formulas in question. Often, truth is defined in terms of (formal) derivability from certain axioms. Frequently, in such circumstances a

more modest claim is made—the claim to truth-in-*S*, where *S* is the particular system in question.

It is worth mentioning here that certain views of truth in arithmetic which claim the Peano axioms to be "analytic" of the concept of number fall under this rubric. Similarly, it is intended to cover conventionalist accounts, since what marks them as conventionalist is the contrast between them and the "realist" account that analyzes (2) by assimilating it to (1) via (3). To make one further distinction, I will *not* automatically call *combinatorial* a view that takes mathematical propositions to be *about* matters combinatorial, either self-referentially or otherwise. For such a view might attempt to analyze mathematical propositions in a "standard" way in terms of the names (if any) they might contain and of the properties they ascribe to the objects within their domain of discourse—which is to say that the underlying concept of truth is essentially Tarski's, though perhaps in so doing they construe the mathematical universe as being populated exclusively by mathematically unorthodox objects: mathematics is merely metamathematics, which is syntax.

It is not my immediate purpose to evaluate these various approaches to the truth of sentences such as (2). I wish here simply to alert us to this distinction—between those views which attribute the obvious syntax (and the obvious semantics) to mathematical statements, and those which ignore the apparent syntax and semantics in order to state truth conditions (or to specify and account for the existing distribution of truth-values) on the basis of more evidently non-semantic syntactic considerations. In the following sections, I will examine both kinds of views in more detail with a particular eye to their relative merits. Suffice it for now to say that I will argue that each kind of account has its merits and defects: each answers to certain interests which we have (or ought to have) in giving an account of mathematical truth. The interests I have in mind are two and these:

(A) Any account of mathematical truth must be recognizably an account of *truth*.

Our account should imply truth conditions for mathematical propositions that are evidently conditions of their *truth*: there must be some general view of truth on the basis of which the property attributed to mathematical propositions when they are said to satisfy the conditions set down by a candidate for an account of truth is indeed truth. I will argue that we have only one such general account, Tarski's, that its essential feature is to define truth in terms of reference (or satisfaction) on the basis of a particular kind of syntactico-semantical analysis of language, and thus that any putative analysis of mathematical truth must be an analysis of a concept which is a truth concept in Tarski's sense. I believe that, suitably elaborated, this condition rules out much: namely all the accounts that I have termed "combinatorial." On the other hand, the account which assimilates (2) above to (1) and (3) obviously meets this condition, as do many variants of it.

My second condition on accounts of mathematical truth presupposes that we have mathematical knowledge, and that such knowledge is no less knowledge for being mathematical. Since our knowledge is of truths, an account of mathematical

truth, to be acceptable, must be consistent with the possibility of having mathematical knowledge: the conditions of the truth of mathematical propositions cannot such as to make it impossible for us to know that they are satisfied. This is not to argue that there cannot be unknowable truths—only that not all truths can be unknowable, for we do know some. The minimal requirement, then, is that a satisfactory account of mathematical truth must be consistent with the possibility that some such truths be knowable. Actually, I will state a stronger requirement: that

(B) Any account of mathematical truth must be useful as part of an explanation of the existence of particular bits of mathematical knowledge.

In short, an acceptable semantics must mesh with an acceptable epistemology.

For example, if I know that Cleveland is between New York and Chicago, it is because there exists a certain relation between the truth conditions for that statement and my present subjective state of belief (whatever may be our accounts of truth and knowledge, they must connect with one another in this way). Similarly, in mathematics, it must be possible to link up what it is for *p* to be true with my knowing that *p*. Though this is extremely vague, I think we can see how condition B tends to rule out accounts which satisfy condition A, and to admit those ruled out by A.[6] For a typical account satisfying A (at least in the case of number theory or set theory) will depict truth conditions in terms of conditions on objects whose nature, as normally conceived, places them beyond the reach of the better understood means of human cognition (e.g., sense perception and the like). The "combinatorial" accounts, on the other hand, usually arise from a sensitivity to *just* this fact and are hence almost always motivated by epistemological reasons. Their virtue lies in providing an account of the nature of mathematical truth based on the procedures we follow in justifying truth claims in mathematics, namely *proof*. It will therefore come as no surprise that *modulo* such an account of mathematical truth, there is little mystery about how we can obtain mathematical knowledge. We need only account for our ability to produce and survey proofs. However, squeezing the balloon at that point apparently makes it bulge on the side of truth: the more nicely we tie up the concep of proof, the more closely we link the definition of proof to combinatorial (rather than semantic) features, the more difficult it is to connect it up with the truth of what is being thus "proved"—or so it would appear.

These then are the two requirements. Separately, they seem innocuous enough. In the balance of this paper I will both defend them further and flesh out the argument that jointly they seem to rule out almost every account of mathematical truth that has been proposed. I will consider in turn the two basic approaches to mathematical truth that I mentioned above, weighing their relative strengths in light of the two fundamental principles that I am advancing. I hope that the principles themselves will receive some illumination and support as I do so.

## 12.3 The Standard View

I will call the "platonistic" account that analyzes (2) as being of the form (3) "the standard view." Its virtues are many, and it is worth enumerating them in some detail before passing to a consideration of its defects.

As I have already pointed out, this account assimilates the logical forms of mathematical propositions to those of apparently similar empirical ones: like other more mundane propositions, mathematical propositions contain predicates, singular terms, quantifiers, etc. What this means is that the truth definitions for individual mathematical theories thus construed will have the same recursion clauses as the ones for their more earthbound brethren. They will not be distinguished in point of logical grammar (provided one doesn't get perverse about non-mathematical statements). That this is a tremendous advantage should be evident, for it means that the logico-grammatical theory we employ in less recondite and more tractable domains will serve us well here. We can do with *one* account and need not invent another for mathematics. This should hold true on virtually any grammatical theory that includes a semantics adequate to account for truth. My bias for Tarski in this context comes from his monumental achievement in giving what is to my knowledge the only viable systematic account of that traditionally elusive concept. Although I will want to differ with Donald Davidson on the subject of truth and meaning in a later section of this paper, let me take this opportunity to endorse fully his urgings[7] that Tarski's account must be extended to natural languages, and that it should be considered at least a necessary condition of an adequate empirical semantics that it provide us with something like a truth definition for the relevant portions of the language.

One consequence of the economy which attends the standard view is that, with truth standardly defined, logical relations are subject to uniform treatment: invariant with subject matter. The same rules of inference may be used and their use accounted for by the same theory which provides us with our ordinary account of consequence. Here too we can dispense with a double standard. If we do not adopt such a view, the inferences permitted in mathematics will need a new, special, account. For example, standard uses of quantifier inferences seem to require for their justification some sort of soundness proof. The formalization of theories in first order logic requires for *its* justification the garantees (provided by the completeness theorem) that all the logical consequences of the postulates will be forthcoming as theorems. Given the standard view, none of these issues present major problems. The obvious answers seem to work. I will defer until we come to combinatorial views any detailed discussion of the problems encountered when the theory of truth is not a standard one. So much for the obvious virtues of this account. What are its faults?

The principal defect of the standard account, when it is not being construed as a strict formalism, is that it violates the second of the two requirements I set down, the one having to do with the integration of an account of mathematical truth into our account of knowledge. In order to make this out with any degree of conviction,

I will have to outline an epistemological picture that I take to be roughly correct and on the basis of which mathematical truths, standardly construed, do not seem to constitute knowledge. I apologize for this detour through the more general problems of epistemology, but I don't see how to make my case without it.

For Hermione to know that the black object she is holding is a truffle is for her (or at least requires her) to be in a certain (perhaps psychological) state.[8] It also requires the cooperation of the world, at least to the extent of permitting the object she is holding to be a truffle. Further—and this is the part on which I would like to lean—in the normal case, that the black object she is holding is a truffle must figure in a causal explanation of her knowing that the black object she is holding is a truffle. On almost any account of explanation, the conditions imposed by the account would insure that in some sense, $p$ must figure in an explanation of Hermione's knowledge that $p$. If what is to be explained is that Hermione knows that $p$, and the explanans is to imply the explanandum (as it must on most accounts), it must also imply $p$, because the explanandum does. So what I require is that $p$ figure in some suitable way in the explanation of Hermione's knowledge. But what is a "suitable way"? One approach which seems promising has been taken recently by Alvin I. Goldman,[9] it involves requiring that Hermione's *believing p* bear a suitable causal relation to the fact, state of affairs, etc. correlated with $p$. One difficulty with this theory is that if we are to spell it out, some account will have to be given of how facts, states of affairs, and/or the like can engage in causal relations. Donald Davidson has recently argued convincingly that the most likely participants in these relations are events, and has in the process given an excellent analysis of how events participate.[10] I am less convinced by his negative thesis excluding other entities than I am by the positive one concerning events. But to fill out my account, I would have to do for facts, state of affairs, etc. what he has done so ably for events. If not, at the very least I would have to argue that with every case of $X$'s knowing that $p$, there is an associated pair of events $e_X$ and $e_p$ which bear some appropriate causal relation to one another and both of which figure in an appropriate way in an explanation of $X$'s knowing that $p$.

Recent philosophic literature contains a spate of articles on the problem of knowledge and related issues (by Harman, Gettier, Skyrms, Unger, Lehrer, and many others).[11] The proffered accounts differ considerably from one another, and it would take me too far afield to comment on the details of each. However, to pick a few, Harman, Goldman, and Skyrms seem to me to be pointing in the general direction in which I think the truth lies. All three are consistent with some causal account of knowledge, and all three seem to be explicating features of the general empiricist view—which for present purposes I should like to espouse, if it is possible to marry such a vague and unspecific bride. (Let me add parenthetically that I don't believe the empiricism I am espousing to be incompatible with our having the linguistic knowledge we have, even if the linguistic knowledge we have is as Chomsky supposes it to be in denying empiricism and affirming some form of rationalism. I don't wish to discuss this here, and I mention it only because the ideas are in the air and ought not to be ignored.) Perhaps one illuminating way of convincing oneself that some such view must be correct is to think of the kinds of

reasons one can offer for claiming that someone *could not* know some particular thing which she claims to know. If we are satisfied that the person in question has normal inferential powers, that the proposition in question is true, etc., then we are thrown back on arguing that the person could not have come into possession of the relevant evidence or reasons: that her four-dimensional space-time worm does not make the necessary (causal) contact with the grounds of the truth of the proposition for her to have the evidence to make the inference (if an inference was relevant). The proposition $p$ places restrictions on what the world can be like. Our knowledge of the world, combined with our understanding of the restrictions placed by $p$ (these are given by the truth conditions of $p$) will often tell us that a given individual could not have come into possession of evidence sufficient to come to know $p$, and we will thus deny her claim to knowledge.

It seems indisputable that the normal account of our knowledge about medium sized objects, in the present, is along the right lines and will involve, causally, some direct reference to the facts known.[12] Furthermore, such knowledge (of houses, trees, truffles, dogs and bread boxes) presents the clearest case and the easiest to deal with. If Hermione knows the language, has normal sensory apparatus, which is functioning normally at the time, etc., etc., then, having extracted her truffle from the can, which is properly labeled, she indeed knows that the black object she is holding is a truffle. And she *know*s it because she can read the label, she can tell a truffle when she sees (and smells) one, because she has read the label, and the object she is holding looks and smells as truffles should, and because it *is* one. Further cases of knowledge can be explained as being based on inferences based on cases such as these. Specifically, what I have in mind is our knowledge of general laws, and, through them, our knowledge of the future and of much of the past. Then will come an account of our knowledge of theories, etc., much along the lines that have been proposed by empiricists, but with the crucial modification introduced by the explicitly causal condition mentioned above—but often left out of modern accounts, largely because these accounts have been at pains to draw careful distinctions between "discovery" and "justification."

In brief, in conjunction with our other knowledge, we use $p$ to determine the range of possible relevant evidence. We use what we know of $X$ (the putative knower) to determine if there could have been an appropriate kind of interaction, if $X$ could have come into possession of enough evidence to warrant her belief that $p$. If not, then she could not know that $p$. The connection between what must be the case if $p$ is true and the causes of $X$'s belief can vary widely. But there is always some connection, and the connection relates the grounds of $X$'s knowledge to the subject matter of $p$.

Such, I think, is the skeleton of an adequate analysis of knowledge. At least, the necessary condition imposed which must belong to such a skeleton is that it must be possible to establish an appropriate sort of connection between the truth conditions of a proposition (as given by the truth definition for the language in which it is embedded) and the grounds on which the proposition is said to be known. In the absence of this, no connection has been established between *having those grounds* and *believing a proposition which is true*. Having those grounds cannot be fitted

into an explanation of *knowing p*. The link between *p* and justifying a belief in *p on those grounds* cannot be made. But if knowledge is properly regarded as justified true belief, then the link *must* be made.

It should come as no surprise that this has been a preamble to pointing out that on this view of knowledge, and on the "standard" view of mathematical truth, it is very difficult to see how mathematical knowledge is possible. For, to start with number theory, if numbers are the kinds of entities they are normally taken to be, then the connection between the truth conditions for the statements of number theory and any relevant events connected with persons who are supposed to have mathematical knowledge cannot be made out. It will be impossible to account for how anyone knows any properly number-theoretical propositions. Our second condition on an account of mathematical truth will not be satisfied, because in order to satisfy it, one must show how the truth conditions for mathematical propositions can be ascertained by us to obtain. One obvious answer—that some of these propositions are true if and only if they are derivable from certain axioms via certain rules—will not help here. For, to be sure, we can ascertain that *those* conditions obtain. But in such a case, what we cannot make out is the link between truth and proof, when truth is directly defined in the standard way. In short, although it may be a truth condition of certain number-theoretic propositions that they be derivable from certain axioms according to certain rules, *that* this is a truth condition must also follow from the account of *truth* if the condition referred to is to help connect truth and knowledge, if it is by their proofs that we know mathematical truths.

Of course, given some set-theoretic account of arithmetic, both the syntax and the semantics of arithmetic can be set out so as superficially to meet the conditions laid down. But the regress that this invites is transparent, for the same questions must then be asked about the set theory in whose terms the answers are couched.

Gödel seems quite aware of the fact that, given a platonistic (i.e. standard) account of mathematical truth, our explanation of how we know the basic postulates must be connected with what we conceive them to mean. Thus, in discussing how we can resolve the continuum problem, once it has been shown to be undecidable by the accepted axioms, he paints the following picture:

> […] the objects of transfinite set theory […] clearly do not belong to the physical world, and even their indirect connection with physical experience is very loose […].
>
> But, despite their remoteness from sense experience, we do have something like the perception also of the objects of set theory, as is seen from the fact that the axioms force themselves upon us as being true. I don't see any reason why we should have less confidence in this kind of perception, i.e. in mathematical intuition, than in sense perception, which induces us to build up physical theories and to expect that future sense perceptions will agree with them, and, moreover, to believe that a question not decidable now has meaning and may be decided in the future.
>
> Gödel [1964] (1990: [271] 267–268)

What is the matter with this picture? Clearly, my own objection is that in the absence of any credible account of how the axioms "force themselves upon us a being true," we should consider the analogy with sense perception and physical

science a misleading one. For what is missing is *precisely* what my second principle demands: an account of the link between our cognitive faculties and the objects of knowledge. In physical science we have at least a start on such an account. We accept as knowledge only those beliefs which we can appropriately relate to our cognitive faculties. To be sure, there is a *superficial* analogy here. For as Gödel points out, we "verify" axioms by deducing consequences from them concerning areas in which we seem to have more direct "perception" (clearer intuitions). But on Gödel's view, we are never told how we know even these, clearer, propositions. For example, the "verifiable" consequences of axioms of higher infinity are (otherwise undecidable) number-theoretical propositions which themselves are "verifiable" by computation up to any given integer. But the story, to be a coherent one anywhere, must tell us how we know statements of computational arithmetic—*if they mean what the standard account would have them mean*. And *that* we are not told. So, the analogy is at best superficial.

If our account of empirical knowledge is an acceptable one, it must be in part because it tries to make the connection evident, in the case of our theoretical knowledge concerning matters not immediately accessible to our senses. This is not to argue for "foundations" of empirical knowledge of any reductionistic sort. Nor is it to suppose that our only justification for empirical beliefs is the "observational" consequences they imply. But it is to try to extract and explain the kernel of truth that has motivated such accounts.

Therefore, in the case of mathematics, in the absence of a coherent account telling how our mathematical intuition is connected with the truth of mathematical propositions, we must conclude that the picture of truth is unsatisfactory. To introduce a speculative historical note, with some foundation in the texts, it might not be unreasonable to suppose that Plato had recourse to the concep of *anamnesis* at least in part to explain how, given the nature of the forms as depicted by him, one could ever have knowledge of them.[13] But this is at best meant to be suggestive of what I mean.

Before passing directly to the combinatorial views, let me mention one possible interpretation of Gödel's and related views which might seem plausible and that I might appear to have overlooked. One might propose to regard the sentences of "computational arithmetic" (variable free sentences) as expressing computational rules, and thus as immediately verifiable by following such rules. Statements with variables, and quantified statements, of arithmetic (and for that matter of set theory) might be seen as instrumental devices for going from "verifiable" statements to "verifiable" statements, on an analogy with instrumentalist views of theories in natural science. Surely something like this was behind Hilbert's views when he regarded quantified statements or quantifiers as "ideal elements." A sophistication might be to regard statements which are "finitely verifiable," whether or not they contained quantifiers, as the basis of mathematics, and the rest as instrumental in simplifying the system of finitely verifiable, or "real" statements. It should be evident that such a view, however attractive, does not count as what I have called a "standard" view. For it does not give a standard interpretation of the quantifiers, nor does it give a standard interpretation of the "real statements"—unless some

metalinguistic version of them can be found which can be fitted into a truth definition for the whole language. Furthermore, the unsophisticated version leaves one wondering why add the rest of mathematics at all, since the "real statements" can be completely and finitely axiomatized. The more sophisticated version, depending as it does on some concept of "finite verifiability" is for that reason open to different construals, with different attendant consequences. But we will return to a discussion of this view almost immediately.

## 12.4 The Combinatorial View

The "combinatorial" view of mathematical truth has its motivation and origins in a realization that, whatever may be the "objects" of mathematics, our knowledge is obtained from proofs. Proofs are or can be (for some, must be)[14] written down or spoken; mathematicians can survey them and come to agree that they *are* proofs. It is on the basis of these proofs that mathematical knowledge is obtained and transmitted. In short, the fact of mathematical knowledge and its (essentially linguistic) means of production and transmission gives their impetus to the class of views I call "combinatorial."

Noting that proofs are seemingly sufficient to produce knowledge, it seeks the grounds of truth in the proofs themselves. This is, of course, not the entire motivation. There is in addition the realization, so central to our last section, that the wildly platonistic picture leaves it a mystery how knowledge can be obtained at all. Given that realization, plus the belief that it is a child of our own begetting (mathematical discovery, on these views, is seldom discovery about an independent reality), it is not surprising that one looks for acts of conception to account for the birth.

To illustrate briefly, one can see in Hilbert's concerns about the infinite[15] some similar motivation. But Hilbert's concern went deeper: he did not simply wish to give a philosophical account of the nature of mathematical knowledge. He feared that unless the account was of a particular sort, the very content and security of mathematics was threatened. He felt that the extension of principles employed in reasoning concerning finite collections to the case of infinite collections needed justification, lest it lead to contradiction, thus putting in question the very applicability of elementary logical principles, such as excluded middle, quantification, etc. The permission is granted because of the simplification in the laws of logic governing the "real," "finitary" statements that the introduction of these "ideal" elements permitted. The logical laws governing "real" statements would otherwise have to be immensely complicated.

The proof of such innocuousness, were it possible to carry it out on Hilbertian principles—if accompanied by a proof of *completeness* of the formal calculus embodying these laws—would permit the substitution of questions of provability for those of truth. Undoubtedly, Hilbert had this in mind. Yet the species of formalism which he advocated (if one may use the definite article for such a holocaust

of indefinite views) considered only the "finitary" statements as being meaningful, the rest as meaningless *but* useful. The point to notice is that in doing this, Hilbert seemed to be making the "knowability" of propositions the touchstone of their meaningfulness and truth. Since, grammatically, it is hard to distinguish between real and ideal statements, to adopt the method of ideal elements *à la* Hilbert is effectively to give up the notion of mathematical truth. If a (syntactically) complete and consistent system could have been constructed, it might have been possible to extend meaning to ideal elements consistently with Hilbert's philosophical principles. But when the theory of proof proved to contain the seeds of its own destruction, the possibility of elaborating a Hilbertian account that meets our two conditions went with it. But we cannot dwell further on Hilbert. His papers are full of enormously stimulating remarks, and a detailed consideration of his views would be endless—for it would contain all the exciting work being done in proof theory, which is stimulated in large part by the Hilbert Program. But I need not sing his praises here. I used him here merely as an illustration.

There is a raft of combinatorial views. I think they may profitably be divided into several groups. It will be evident from the divisions and the views contained in them that in certain cases it is even *prima facie* a misnomer to call some of them views of mathematical *truth*. They don't pretend that what they define is truth; indeed, for some there is no such animal. It will be the burden of this section to argue that on none of these views does one get an adequate account of mathematical truth: that the concepts defined, even by those who think they are getting at truth, are not concepts of *truth* at all. This would require, for its full documentation, a certain analysis of truth which I cannot conceivably give here—any more than it was possible to provide the analysis of knowledge needed in the preceding section. So, as you hunger after knowledge, so shall ye thirst for truth. But where it seems useful, I will try to make some comments pointing in the direction I would take. I will now enumerate and discuss the kinds of views I am lumping together under the "combinatorial" umbrella.

(a) *Crude Formalism*: Mathematics consists of the manipulation of meaningless marks according to certain rules. The "propositions" are neither true nor false. Meaningful questions are limited to combinatorial ones concerning the possible results of applying rules (e.g. syntactic consistency, whether this or that formula is derivable, etc.). Though it is doubtful that Hilbert ever held such a view, he certainly came very close in certain instances. And, as I suggested above, it is not unreasonable to suppose that the Hilbert Program for the analysis of mathematical proof was pointed in the direction of such a view: if its goals could have been achieved, the reduction claimed to exist by crude formalists would have been considerably more palatable, for no mathematics need have been relinquished, only some interpretation of it. The consequent epistemological gain from the reduction of the concept of mathematical proof to that of derivability within a single, provably complete and consistent formal system would make almost any philosophical price worth paying. Such a view is clearly less an analysis of truth in mathematics than an open refusal to speak

in such terms. It therefore clearly fails to satisfy my first condition, and just as clearly meets the second. I will not return to this view.

(b) *Mathematics is the "if... then" science*: it consists of deriving logical consequences from axioms which themselves have no truth-value. We simply investigate the deductive properties of axiom systems, asserting sentences of the form

If *Axioms* then *Theorem*.

This view should sound familiar. For, substitute first order derivability for logical implication and you get one brand of Formalism right back. This view is motivated by a denial that mathematical axioms are true or false and that mathematical proof eventuates in theorems that are detached from the axioms. It falls short of the Crude Formalism described under (a) by claiming that there are at least some candidates for assertability which are not strictly combinatorial claims: the "if… then" statements, or at least so it would seem. The view splits off from Formalism decisively if its proponents are willing to deny that there is any adequate "formal" explication of logical consequence. For those among you (us) whose early toilet training included the ritual recitation that logical consequence is adequately captured by first order derivability, this might seem difficult to conceive. But as early as 1933 and 1936,[16] Alfred Tarski discussed concepts of logical consequence which were demonstrably (after 1931) non formal, for they considered that for systems of arithmetic, [$(x)$ $Fx$] was a consequence of the set of sentences resulting from [$Fy$] by replacing "$y$" with each standard numeral of the system. Since, given any consistent formalization of number theory, there will be predicates $F$ for which all the relevant numerical instances will be theorems but for which the universally quantified statement will not, logical consequence thus conceived is unformalizable. There are, accordingly, two species of "if… then" theorists: those who are satisfied with first order derivability as an equivalent of logical consequence, and those who are not. Among those who *are* satisfied with it the position reduces to a kind of Formalism if they do not base their satisfaction on the completeness proof for first order logic.

By and large, I fail to understand what concept of logical consequence those who don't accept the completeness proof may have in mind. Though there seem to be partial motivations through Herbrand-Gentzen-Ackermann procedures, I have not found them completely satisfactory. In any event, I consider views based on these theorems intermediate between the flat refusal to *discuss* soundness and consistency, claiming first order derivability as the explanation of logical consequence, and the full-fledged acceptance of the completeness proof in all its infinitistic glory.

Those who do accept the completeness proof must accept some standard concept of first order validity which the completeness proof shows to be coincident with derivability (*modulo* soundness as well, of course). This will invariably be some set-theoretic concept, usually truth in every model. But then, the view cannot be applied to set theory as well, and there is some mathematics (set theory) which is not simply an "if… then" discipline and for which some different account must be given: either a combinatorial one or a platonistic one. In either case I need not

consider them further in this context. The same will be true of the "if… then" theorists who are not satisfied with first order derivability as an equivalent for their concept of logical consequence. For them there will be all the more reason to depend on some set-theoretic concept—which, after all, is all we have. Until we get a better explication of logical consequence than truth of the conclusion in every model of the premises it is unlikely that the "if… then" view will approach adequacy as an explanation of or substitute for mathematical truth. In any case, it should be clear that my general argument applies to this class of views.

One further problem with this view follows from matters pointed out by Tarski in 1936. In explicating logical consequence he pointed out that his analysis depended on the notion of a model, which in turn depended on the concept of a *logical constant*—for which we had *and have* no explication—only a more or less adequate list. Hence, the concepts of logical consequence, logical truth, etc. will be relative to the choice of constants. One might think of the choice of constants as a third dogma, shared not only by empiricists but by quasi-, neo- and proto-pragmatists as well.

(c) *Mathematical propositions (when true) are analytically true*, i.e. true in virtue of meanings alone. Reluctant as I am to reach into this Pandora's box, I must. For its content cannot be dismissed out of hand. This currently much maligned group of views dominated the empiricist scene for a long time, and has only recently found its much deserved disfavor, largely as a result of Quine's attacks on the notions of meaning on which they rest.

*Logicism.* Mathematics is reducible to logic. Each mathematical concept is definable in terms of concepts belonging to logic, and, under these definitions, the theorems of mathematics are translatable into theorems of logic. Furthermore, these definitions accurately render the meanings of the mathematical terms. These claims, plus the claim that the propositions of logic are *themselves* analytic—that they are true in virtue of the meaning of *their* constituent terms—amount to the classical statement of the logicist position.

There are many variations under this same rubric but I will not play them all. I am calling "logicism" that class of views which conclude to the analyticity of mathematical propositions on the basis of the *explicit definability* of the concepts of mathematics in terms of those of logic. For our purposes, the further classification that will be useful is on the basis of how those who hold this position view logical truth: for, even granting the enormously moot claim of the definability of the mathematical concepts in terms of those of logic, the view requires for its support the additional step providing for the analyticity of the laws of logic to which mathematics is allegedly thus reducible. For if *they* cannot be shown to be analytic in the relevant sense, the view collapses: the first step in the argument having been to argue that the definition showed mathematical propositions to be analytically *equivalent* with those of logic. Therefore, by the first step of the argument, mathematical propositions are analytic only if those to which they are "reducible" are.

We can conclude that logicism would fall with the collapse of the translatability claim.

But if that claim withstands Quine's assault on it, additional buttressing will be needed in the form of a suitable analysis of truth in logic. Let us bypass translatability and address ourselves to logical truth, granting for the sake of argument that the logicist's definitions reflect the necessary synonymies.

If mathematics is reducible to logic, then the logic to which mathematics is reducible is set theory, type theory, or some such. We can limit our attention to set theory as the "logic" needed to complete the logicist account of mathematical truth. Type theory would do as a possible alternative but would present no essentially different solutions—except possibly under Russell's "no class" interpretation of *Principia Mathematica*. But this view, appealing though it be, will not serve the logicist who wishes to preserve classical mathematics. For with logic so construed, the first part of his argument collapses. Mathematics is not reducible to logic.

This leaves us with essentially three positions with which to deal. We have already considered the first: it is the standard realistic account of set-theoretic truth, the account to which we saw Gödel subscribing some pages back. It does not belong in this section. The second is the view, mentioned above, that the axioms and rules of logic constitute implicit definitions of the undefined terms occurring in them, where it is supposed that the truth of the axioms (and theorems) is guaranteed by the meanings of the terms. Finally, the third view is that truth (and meaning) are conferred upon these axioms by explicit convention/stipulation—i.e. that these axioms and rules represent conventions that we stipulate for the use of these expressions, in virtue of which the sentences containing them are true.

Both of these latter views have non-logistic counterparts when what is at issue is mathematical truth with no claim for the reducibility of mathematics to logic. For example, both have been advanced as accounts of *arithmetic* truth. It should be fairly clear from what follows how most of the objections I will raise to these conceptions of *logical* truth apply, *mutatis mutandis*, to the non-logicist theories of mathematical truth. This should make it unnecessary to treat these other views separately.

*Truth by convention*. Quine, in his classic paper on this subject (Quine [1935] 1976), dealt clearly, convincingly and decisively with the view that the truths of logic are to be accounted for as the products of convention—far better than I could hope to do here. He pointed out that, since there are infinitely many truths to be accounted for, the characterization of the eligible sentences as truths must be wholesale rather than retail. But finitely expressed wholesale characterization can only come by means of general principles—and if we are supposed not to understand any logic at all, we cannot extract the individual instances from the general principles: we would need logic for such a task.

Convincing as this may be, I wish to add another argument—not because I think this dead horse needs further flogging, but because the principal point I wish to make is insufficiently clearly made in Quine's argument. Indeed, Quine grants the conventionalist certain principles I should like to deny him. In resting his case against conventionalism on the need for a wholesale characterization of infinitely

many truths, Quine concedes that if there were only finitely many truths to be accounted for, the conventionalist might make out his case. He says:

> If the truth assignments were made one by one, rather than an infinite number at a time, the above difficulty would disappear; truths of logic […] would simply be asserted severally by fiat, and the problem of inferring them from more general conventions would not arise.
>
> Quine [1935] (1976: [344] 105)

It appears that on his view if some finitistic way could be found to make sentences of logic wear their truth-values upon their sleeves, the objections to this account of truth would disappear—for we would have determined truth-values for all the sentences, which is all that one could ask.

I wonder, however, what it is supposed such a distribution of the word "true" would accomplish. It surely cannot suffice in order to determine a concept of truth to assign "truth"-values to each and every sentence of the language. Suppose now that the language is set theory, in some first order formalization, and let those sentences with an even number of horseshoes be "true." What would make such an assignment of the predicate "true" the determination of the *concept of truth*? Surely not simply the use of that time honored monosyllable. Tarski has supplied us with Convention T as a necessary and sufficient condition on a definition of truth for a particular language (Tarski [1933] 1983). A mere distribution of truth-values will not suffice to satisfy Tarski's criterion. For a mere distribution of truth-values will not provide us with the sentence-by-sentence translations needed for the satisfaction of Convention T. I submit that what would be missing—hard as it is to state—is the theoretical apparatus employed by Tarski in providing his truth definitions. I submit first that we would not accept as a concept of truth something that failed to satisfy Convention T, but also, something that, though it satisfied Convention T, did not define truth by analyzing the language as having a structure involving predication, naming, quantification, etc. In short, I feel that a definition of truth which does not proceed by the customary recursion clauses for the customary grammatical forms may not be deemed adequate—even if it satisfied Convention T.

This brings me to the difference I would raise between my views and those of Donald Davidson's, and which I think is relevant to the case at hand. I should like to invert Davidson's argument of "Truth and Meaning." He claims essentially that the grasp we have on the concept of meaning (and presumably reference as well) is through the concept of truth—and that to have apparatus sufficient for the satisfaction of Convention T is for all practical purposes to have enough for the determination of meaning. The Quine of "Truth by Convention" felt that to determine the truth-values of all the contexts which contain a word suffices to determine its meaning. I should like to suggest, however, that our concept of truth, insofar as we have one, proceeds through the mediation of the concepts Tarski has used to define it for the class of languages he has considered—that the essence of Tarski's contribution goes farther than Convention T but includes the schemata for the actual definition as well: that an analysis of truth for a language that did not proceed through the familiar devices of predication, quantification, etc., might not prove satisfactory.

If this is all near the mark, then it should be clear why "combinatorial" views of the nature of mathematical truth fail on my account. They avoid what seems to me to be the necessary route to an account of truth: through the subject matter of the sentences whose truth is being defined. Being motivated by epistemological considerations, they come up with truth conditions whose satisfaction or non-satisfaction it is relatively possible for mere mortals to ascertain, but they pay the price of being unable to connect these so-called "truth conditions" with the truth of the propositions for which they are conditions.

So, for Quinean reasons, I think it is impossible to make out a satisfactory doctrine of logical-truth-by-convention. If for some reason it be granted that the truths of first order logic do not stem from conventions, then the problem of accounting for mathematical truth splits into two: first order logic and the rest. Ignoring logic, someone may still wish to claim that the rest (set theory, for logicists; set theory, number theory, and other things for non-logicists) consists of conventions formalized in first order logic. This view, I think, falls on my general objection that such a concept of convention need not bring *truth* along with it. Indeed it is clear that it does not. For, ignoring my other more general objections, once the logic is fixed, no matter how, it becomes possible that the conventions thus stipulated turn out to be inconsistent relative to the interpretation of the logic, which is already fixed. Hence it is impossible to maintain the claim that setting down conventions *guarantees* truth. But if it does not *guarantee* truth, what distinguishes those cases in which it provides it from those in which it does not? Consistency cannot be the answer. To adopt it as such is to *mis*construe the significance of the fact that *in*consistency is *proof* that truth has not been attained. The deeper reason is that stipulation makes no connection between the sentences and their subject matter —stipulation does not provide truth definitions in the sense in which we have been led to expect them.

As for the final view we shall consider: that the postulates are *implicit* definitions of existing concepts (as opposed to stipulations for how new ones are to be understood), very much the same objection can be raised, if it is meant as an explanation of how we know the axioms to be true (we learned the language by learning *these* postulates). Otherwise, such a view can be taken simply as stating that the truths of the field can be conveniently regimented in this or that axiomatic form. When thus taken, it is innocuous. But it is also totally devoid of explanatory content on the point at issue: the grounds of mathematical truth. Yet it is hard to see how else to construe it, if one views as gratuitous the "explanation" that the meanings of the logical and mathematical primitives "dictate" the truth of the postulates.

To clarify the point, consider Russell's oft cited dictum: "The method of 'postulating' what we want has many advantages: they are the same as the advantages of theft over honest toil" (Russell 1919: 71). On my view, that's false. For with theft you at least come away with the loot (not that crime pays—only if you don't get it, it isn't theft). Implicit definition however, is incapable of bringing truth. It is practically, as well as morally deficient.

There are other positions, but I won't bore you further by enumerating and discussing them. I hope that these remarks will suffice to make the general point of this section clear: Truth is truth *à la* Tarski. Truth *à la* Tarski is satisfaction of Convention T *in a particular way* (i.e. via satisfaction, reference, quantification, etc.). "Combinatorial" views don't give us truth—though the "truth conditions" they assign to sentences can be known to obtain without having to commune directly with the Forms.

## 12.5 Conclusions

I will conclude shortly and briefly by answering a few objections carefully chosen at random from those that might immediately leap to mind. I hope they will help summarize the views and stave off the day of reckoning:

*Objection 1*. Don't the same objections you raise against platonistic theories apply to combinatorial ones as well?

*Objection 2*. You have stacked the deck.

*Objection 3*. Your position, depending as it does on being able to segregate bits of experience as being relevant or irrelevant to the truth of propositions one-by-one is incompatible with a totally holistic epistemology, for which your problems don't even arise.

*Objection 4*. What can your condition A (and its attendant justification) amount to in light of the fact that on Tarski's own theory, you cannot conceivably elaborate an analysis of mathematical truth which eventuates in a definition? You are citing Tarski's theory of truth—which is a theory concerning *truth definitions*—in support of the requirement on an analysis which could not eventuate in a definition. What is the relevance of Tarski?[17]

*Answer to objection 1: Combinatorial views*

It may not be immediately evident *how* we have any more contact with the universals required for the proper functioning of combinatorial views than we do with their more abstract brethren summoned up by the platonistic positions—say, for example, the Zermelo-Fraenkel sets. Thus it may seem that the picture is even darker than I have painted it—combinatorial views are doubly bad: not only do they not give us concepts of truth, but they also fail to provide "truth conditions" which are any more knowable than those of the platonist.

As I would tell my six-year-old son, that's a long story. Briefly, I think the answer is no. I think we have a fair idea that *non* universals would suffice to constitute a proof. This is the virtue of formalization: it provides us with a regimentation of proofs—a standard form for proofs. Of course this is not to say that the concept of proof is entirely—or even almost—clear. It is only to point out that we have reasonably statable sufficient conditions on material objects which we can, on occasion, perceptually recognize to have been met. This was the key to Hilbert's

insight: our truck is with finite objects, always: no matter how infinite may be the apparent object of our thought. So, he felt, in some sense it must be possible to explain the infinite through the finite.

*Answer to objection 2: Truth versus knowledge*

I have spoken all along as if it were our concept of mathematical truth that contained the deficiency. Clearly, one could with equal justice suppose the difficulty to stem form our concept of knowledge. One more cautious still would simply point to the insufficiency of these concepts as we understand them for answering the questions we want answered. To saddle truth with the entire responsibility is perhaps a more dramatic way of pointing to the inadequacy of our *total* account. It is also to indicate what I think is true but cannot argue for: that the account of knowledge I sketched is correct, and it is mathematical truth that cannot bear the burden. Such, at least, is my intuition.

I make these remarks to anticipate an obvious objection: didn't I stack the deck in imposing my causal condition on knowledge? No. I did not. My claim is that with the concepts of knowledge and truth explicated as I have suggested, we do not seem to have adequate accounts of mathematical truth and mathematical knowledge. I am open to suggestion on how the analysis of either concept might be improved to remedy this defect. This is simply a bit of philosophical anthropology: I did *not* choose my analyses with a view to supporting my conclusion. They just do, I hope.

*Answer to objection 3: Argument against total holism*

The presumption is that for each proposition, certain obtainable conditions are *sufficient* to produce knowledge—and that these conditions are connected *in some way* with the content of the proposition, as spelled out by a truth definition for the language. Now, for abstruse physical theories, much complication obtains—the range of relevant experiences is very broad, etc. For most statements about breadbox sized objects, it is relatively narrow. For mathematics—seemingly more remote than even the most abstruse physical theory—the startling thing is that *proof* is sufficient to produce knowledge. Hence it must be sufficient for truth: but how? If you like—the problem of mathematical truth is to account for how *proof* produces *knowledge*. Its connection with *us* is evident enough—but what is its connection with truth? Holist views, satisfactory as they may be on other grounds, blur the very fact that I think needs explaining by telling a fairy tale which jumbles up all kinds of propositions with a net which impinges on experience at the well worn edges.

*Answer to objection 4: Truth*

Tarski's theory does not provide *a* truth definition. It provides conditions on truth definitions. I have suggested that, *implicitly*, it provides more than the *explicit* requirements it lays down—that Tarski's account is valuable not only for containing Convention T but also and perhaps principally for showing *how* Convention T can be satisfied in individual cases. In brief, I think there is a concept

of Truth. Tarski has shown that such a concept, if consistent, must elude adequate definitions—that only concepts of truth in specific languages can be defined.

One has two options here: conclude that there is no general concept of truth, or of truth in mathematics (without metalinguistic regress)—that no theory is possible which accounts for why truth-in-*L* is truth, that truth is a hopelessly relative concept. Or one could seek for some kind of "absolute" account. Gödel makes this "absolute"/"relative" distinction in his 1946 lecture at Princeton, but somewhat obscurely.[18] I opt for the latter course and my requirement amounts to a claim that Tarski's theory contains implicitly the germ of such a theory.

## Notes

1. The Catastrophe Theory was so baptized in a symposium entitled "Is Any Set Theory True?" at a joint meeting of the American Philosophical Association and the ASL, December 27, 1967 in Boston. The Symposiasts, Joseph S. Ullian, Donald Martin and Saul Kripke, were unanimous in decrying the Catastrophe Theory, and W. V. O. Quine was unanimous in upholding a historical version of it.
2. One might think it possible to distinguish two different kinds of propositions independent of any given set of first order axioms for set theory: (a) those independent for Gödelian reasons—i.e. whose undecidability stems from Gödelian incompleteness, and (b) those independent for set-theoretic reasons—presumably the rest. If such a distinction were viable, one believing in the "implicit definition" view might think that propositions independent under (a) have a truth-value, while those independent under (b) don't and represent real gaps in our concept of set. But, for reasons which would take us far afield, it seems that such distinction could not be made out to do the needed work.
3. I do not insist on a privileged *direction* of possible illumination, for often it is only incompleteness (or independence) results that force us to look at concepts of truth which we could *at best* be said to have held *implicitly* prior to being awakened from our dogmatic slumbers. It suffices for my small point (which is but illustrative) to recognize the importance of the connection, regardless of which is the chicken and which the egg.
4. Let us ignore the complication introduced by the fact that a large city is not something large and a city, but more like something large for a city.
5. See Benacerraf (1965). *Editor's note.*
6. I see possible exceptions: for example, the class of views on which all of mathematics is metamathematics and on which every mathematical sentence receives an interpretation via a truth definition. Views on which mathematics consists simply in turning a generative crank on a black box that prints out meaningless symbols are not even in the ballpark we are considering, for (2) above would, on such views, either not be a mathematical statement, or would, at any rate, lack a truth-value.
7. See, for example, Davidson [1967a] (1984). Davidson makes in this article some rather extravagant claims for a Tarskian theory of truth with which I do not wish here to associate myself: specifically, "if [some sentence of the form

"*s* is true if and only if *p*"] followed from a characterization of the predicate "is true" that led to the invariable pairing of truths with truths and falsehoods with falsehoods—then there would not, I think, be anything essential to the idea of meaning that remained to be captured" (Davidson op. cit.: [312] 26).
This much zeal I do not have, for it is easy to see that for an infinite language for which there is at least one such theory, there are indefinitely many different ones with paired sentences as different in meaning as you like (consistent with likeness of truth-value). Though this would not trouble one, like Quine, who shuns even likeness of truth-value as a condition in translation, I must confess it troubles me. Given Quine's radical stand, I can hardly regard this objection as a *refutation* of Davidson's view, but I think it relevant to register some form of mild protest, to stand up and be counted, as it were.

8. If possible, I would like to avoid taking any stand on the host of issues in the philosophy of mind or psychology concerning the nature of psychological states. Any view on which Hermione can learn that the cat is on the mat by looking at a real cat on a real mat will do for my purposes. If looking at a cat on a mat puts Hermione into a state and you wish to call that state a physical, or psychological, or even physiological state, I will not object—so long as it is understood that such a state, if it is her state of knowledge, is causally related in an appropriate way to the cat's having been on the mat when she looked. If there is no such state, then so much the worse for my view.
9. See Goldman (1967). There are, I fear, difficulties with Goldman's account, particularly with those cases in which Hermione's belief that *p* and the fact, state of affairs, event, corresponding to *p* stem from a common cause and it is their relation to that cause that is (in part) responsible for Hermione's belief constituting knowledge.
10. See, e.g., Davidson [1967b] (1980), [1969] (1980) on the role of events in causal explanations. *Editor's note.*
11. See, e.g., the discussions in Gettier (1963), Harman (1973), Lehrer and Paxson (1969), Skyrms (1967), and Unger (1968). *Editor's note*.
12. For an illuminating account of perception, see Grice [1961] (1989). But of course, this is at best a small part of the problem.
13. "The soul, then, as being immortal, and having been born again many times, and having seen all things that exist, whether in this world or in the world below, has knowledge of them all " (Plato [Approx. 402 BC] 1949: 81).
14. Intuitionists, on the other hand, consider the linguistic expressibility of constructions and proofs an incidental feature of mathematics, whose subject matter is "mental constructions." This and other intuitionistic views makes it difficult to deal with the position in the context of this paper and I fear I will have to bypass discussion of it. For an excellent account, see Kreisel (1962).
15. Hilbert advocated the "extension" of these principles to reasoning about infinite domains, provided that it could be shown that such extensions did not lead to contradiction. Referring to statements which involved essential reference to the finite or "real" statements, he advocated thinking of the rest as introduced in the

manner of the "ideal" elements of projective geometry, to simplify laws concerning the real statements.

16. See Tarski (1936), [1936] (1983). *Editor's note*.
17. I am indebted, if that is the right word, to Bill Tait for this objection.
18. Gödel does make a distinction between relative and absolute notions of demonstrability and definability (and of ordinal definability in particular) in Gödel [1946] (1990). His remarks are prompted by Tarksi's stress, in his talk at the Princeton University bicentennial conference on *Problems of mathematics* in December 1946, "on the great importance of the concept of general recursiveness (or Turing computability)" (Gödel op. cit.: [1] 150).
    Gödel never reacted explicitely to the German version of Tarski's 1933 article on the concept of truth in formalized languages. I am indebted to Mark van Atten for this information. See Tarski (1933), [1933] (1983) in the References below. I am indebted to Philippe de Rouilhan for details pertaining to the successive editions of Tarski (1933). *Editor's note*.

## References


Cohen, P. J. (1966). *Set theory and the continuum hypothesis*. New York: W. A. Benjamin Inc.

Davidson, D. [1967a] (1984). Truth and meaning. In *Inquiries into truth and interpretation* (pp. 17–36). Oxford: Clarendon Press.

Davidson, D. [1967b] (1980). Causal relations. In *Essays on actions and events* (pp. 149–162). Oxford: Clarendon Press.

Davidson, D. [1969] (1980). The individuation of events. In *Essays on actions and events* (pp. 163–180). Oxford: Clarendon Press.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123.

Gödel, K. [1946] (1990). Remarks before the Princeton bicentennial conference on problems in mathematics. In S. Feferman (Ed.), *Collected works* (Vol. II: *Publications 1938–1974,* pp. [1–4] 150–153). Oxford: Oxford UP.

Gödel, K. [1964] (1990). What is Cantor's continuum problem? [Revised and expanded version of Gödel [1947] 1990]. In S. Feferman (Ed.), *Collected works* (Vol. II: *Publications 1938–1974,* pp. [259–273] 254–270). Oxford: Oxford UP.

Goldman, A. I. (1967). A causal theory of knowing. *The Journal of Philosophy, 64*(12), 357–372.

Grice, H. P. [1961] (1989). The causal theory of perception. In *Studies in the way of words* (pp. 224–247). Cambridge, Mass. Harvard UP.

Harman, G. (1973). *Thought*. Princeton, NJ: Princeton UP.

Kreisel, G. (1962). Foundation of intuitionistic logic. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science—Proceedings of the 1960 International Congress* (pp. 198–210). Stanford: Stanford UP.

Lehrer, K. & Paxson, T., Jr. (1969). Knowledge: Undefeated justified true belief. *The Journal of Philosophy, 66*(8), 225–237.

Plato ([Approx. 402 BC] 1949). *Meno [Μένων]* (translated from the Greek by B. Jowett). New York: Liberal Arts Press; Indianapolis: Bobbs-Merrill.

Russell, B. (1919). *Introduction to mathematical philosophy*. London: George Allen and Unwin.

Skyrms, B. (1967). The explication of '*X* knows that *p*'. *The Journal of Philosophy, 64*(12), 373–389.

Tarski, A. (1933). *Projecie prawdy w jezykach nauk dedukcyjnych* [The concept of truth in the languages of deductive sciences]. Warszawa [German ed. as "Der Wahrheitsbegriff in den formalisierten Sprachen" by L. Blaustein, *Studia Philosophica* (Lemberg), Vol. 1, 1935, pp. 261–405; English ed. as "The Concept of Truth in Formalized Languages" by J. H. Woodger in Tarski [1933] 1983].

Tarski, A. [1933] (1983). The concept of truth in formalized languages [English edition of "Der Warheitsbegriff in den formalisierten Sprachen"]. In J. H. Woodger (Ed.), *Logic, semantics, metamathematics: Papers from 1923 to 1938* (1956); second edition by J. Corcoran (pp. 152–178). Indianapolis: Hackett Publishing Company.

Tarski, A. (1936). Über den Begriff der logischen Folgerung. In *Actes du Congrès International de Philosophie Scientifique,* Sorbonne, Paris, 1935 (Vol. 7, pp. 1–11), *Actualités Scientifiques et Industrielles* (Vol. 394), Hermann, Paris, 1936 [translated into English as "On the Concept of Logical Consequence" in Tarski [1936] 1983].

Tarski, A. [1936] (1983). On the concept of logical consequence [English translation of "Über den Begriff der logischen Folgerung"]. In J. H. Woodger (1956) *Logic, semantics, metamathematics: Papers from 1923 to 1938*; second edition by J. Corcoran (pp. 409–420). Indianapolis: Hackett Publishing Company.

Unger, P. (1968). An analysis of factual knowledge. *The Journal of Philosophy, 65*(6), 157–170.

Quine, W. V. O. [1935] (1976). Truth by convention. In *The ways of paradox and other essays* (pp. 77–106) [Revised and enlarged edition]. Cambridge, Mass. Harvard UP.

# Chapter 13
# Comments on Reduction (Lecture in a Graduate Seminar ~1975)

**Paul Benacerraf**

There appear to be two kinds of theories on the nature of numbers, distinguished by the way they interpret their results. I am speaking, of course, of theories which conclude, or appear to conclude that, e.g., numbers are sets—that, say, 3 is some particular set. There are, on the one hand, theories which I will refer to as *realist* theories; and there are Quinean theories, which we may think of as *instrumentalist* theories. As with most things in this murky area, the distinction is hard to make out, but also, I think, important. We may be forced to abandon it in the end, but if so, much will have come tumbling down with it.

Theories of the first sort conclude, at the end of some argument, that numbers are sets, properties, extensions of concepts, or what have you. Most commonly, they do so on the basis of linguistic evidence—on the basis of an examination of the meaning of the linguistic apparatus employed in number talk, by semantic descent to the material mode; they draw conclusions about the numbers from premises about the language we employ in seeming to talk (or indeed in talking) about numbers. Such, I think, is Frege's case. His *motivation* was epistemological; he wanted to find the roots of *our knowledge* of numbers. But his route to those roots was first of all to question the nature of the objects of which we had supposed knowledge, and then to proceed to answer his questions by a trip through the language, to an analysis of its meaning, and ultimately, he thought, eventuating in the reference, with a clearer picture of the objects themselves, only a *part* of the nature of which was revealed in arithmetic talk. This seems to me the only

The paper was written for a graduate seminar in the mid 1970s, devoted to the philosophical issues addressed in Benacerraf (1965). Benacerraf spent 1979–1980 at the Palo Alto Center for the Behavioral Sciences, where he wrote his Skolem paper (Benacerraf 1985). Although it is uneasy to date the typescript with precision, from what Benacerraf recollects, the paper must have been written before his stay in Palo Alto.

Minor revisions of style have been made by Paul Benacerraf in 2015. *Editor's note.*

P. Benacerraf (✉)
Princeton University, Princeton, USA
e-mail: paulbena@princeton.edu



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6_13

hypothesis that makes any sense of Frege's claim that his investigation shows the propositions of arithmetic to be analytic, even on his expanded sense: definitional contractions of certain laws of logic. He aimed at accounting for mathematical knowledge by showing how it was a case of *logical* knowledge, and he realized better than anyone that to claim identity of reference for the two vocabularies was simply not enough: on Frege's theory of these matters, that would not have yielded the desired *epistemological* conclusion. He had to show that we *got at* the reference in similar ways in both cases: that the mode of presentation of the reference permitted us to know the objects being referred to (the numbers) through their presentations as extensions of concepts. This would, of course, satisfy only a necessary condition of the analyticity of the mathematical laws. A further account would have to be provided which explained the grounds of the truth of the logical laws to which the reduction was effected. This might have been done by thinking of logical knowledge as part and parcel of speaking the language (in a linguistic version perhaps alien to Frege), or of thinking (in a semantical version which would sit better with him). In each case, what is lacking is a clear specification of the behavioral component. In the linguistic case, knowledge that $p$ is reduced to knowledge *how* to use the language, and must be unpacked in terms of its notoriously slippery concepts—concepts incapable of securing the epistemological advantages required for analytic propositions. In the case where what is at issue is not the linguistic version, but the conceptual one (logic is "the laws of thought"), the tie to behavior is even more tenuous and less able to do the required job. For Frege's aversion to psychologism in logic cut the cord that for others might have tied the theory of meaning to thought and behavior. I mention this in passing if only to note that what promise Frege's theory has of explaining our *knowledge* is undercut by his reluctance to speak of our *belief*. But even if we believe that he will ultimately fail—that another necessary condition for an ultimately satisfactory account will never materialize, it might prove instructive to pursue his argument as far as it will take us.

So, to return to it, we just noted that to establish identity of reference would be insufficient for Frege. He needed to establish something more like identity or similarity of mode of presentation of reference: something akin to sense. *That* explains why the accounts in sections I and II of my paper[1] would not be satisfactory Fregean accounts. Although they seem to me to be perfectly adequate as such accounts go (they fulfil all the conditions I singled out as necessary), there seems to be no way to choose between Ernie's and Johnny's numbers on the basis of our knowledge of number theory. The referent of "3" has not sufficiently clearly been shown to be connected with the referent of "$(\Lambda)$)," or with its von Neumann surrogate. Frege, who defined the numbers directly as the cardinals—i.e. who built the notion of cardinality *into the very definition of the numbers*—thought he had thereby found the route to the roots of our arithmetic knowledge. It is *this* presumption that gives the Frege-Russell numbers first priority on correctness, or so it seems to me. But, if I am right about all this, then neither does it suffice to yield a victory for the Frege-Russell numbers. For the following account will do *just as well* and meets anything that can be dredged up from the above comments as an

additional condition which Frege's account meets but which Ernie's and Johnny's do not. In some set theory (or type theory—the modifications to be made are obvious), pick out a sequence **Z** (says Zermelo's) and define "$S_z$," the concept of immediate successor in **Z**, the least element of which we will call L. In terms of $S_z$ and L, define "$<_z$" and then give contextual definitions of cardinality as follows:

$$C(a, L) =_{df} (\exists w)\,(w \text{ maps a } 1-1 \text{ onto} \{y{:}y <_z L\})$$
$$C(a, S_z x) =_{df} (\exists b)\,(b \in a \,\&\, (\{c|c \in a \,\&\, c \neq b\}, x)).$$

Thus, if one of the elements of $S_z$ is Julius Caesar, this defines what it is for a set to be of the cardinality Julius Caesar. It only remains now to define the numbers as follows:

$$Nn =_{df} \{x{:}\ (\exists y)\, C(y, x)\}$$

and, more specifically,

$$0 =_{df} L$$
$$S_{Nn} =_{df} S_z x.$$

Since, in an ideal world, C induces equivalence classes, i.e. divides the universe into classes of equinumerous classes, it is a fine candidate for the relation of cardinality. The numbers (Nn) are then satisfactorily introduced as those objects which, through their association via C with classes, represent the cardinality of classes. It could hardly be clearer that these are fine and worthy candidates for numbers. By the time we are told that the elements of $S_z$ are the numbers, we have been amply prepared by noting their essential involvement in the concept of cardinality, which is after all the foundation of the concept of number. The edge that the Frege-Russell numbers seemed to hold over Ernie's and Johnny's is not one which they hold over the numbers so defined. Ernie's and Johnny's numbers seem like a put up job, *es post facto*. Who, looking at their definitions but unaware of the *arrière pensée* in the mind of Zermelo or von Neumann would have recognized them as numbers? Not so with these definitions, however. When the definitions of Nn is produced, it is clear at once that it is the numbers that have been introduced.

For the above argument to be made out in any detail on Frege's behalf, a lot of things have to go right with the theory of meaning, and more particularly with the theory of sense and reference. But even if it *is* right, if it *can* be made to work, it fails in its ultimate purpose, which was to find the genuine numbers and not make do with ersatz surrogates. For even if the present *account* were superior to the ones handed Ernie and Johnny, it is clearly consistent with the identification of the numbers as *any* sequence, names for whose members we could generate in a manner suitable for deployment in the definition of cardinality. Reflection upon this fact leads to the following disjunctive conclusion for the Fregean entreprise.

Either (i) there are facts about numbers which we have yet to uncover *as* being about numbers (or uncover at all) and which discriminate which ω-sequence is really the numbers—and this independently of whether or not such facts are discoverable, or even, if discovered, are of such a kind as to warrant including them as necessary conditions on a correct analysis of the concept of number (that, after all, is more a comment about what correct analyses should do than about numbers), or (ii) there are no such facts, and we should rethink the quest on which we found Frege embarking—the quest for the True Numbers.

In this latter vein, the simplest and most immediate conclusion we might be inclined to draw is that the advantage claimed for the Frege-Russell numbers and perhaps as well for the account of the present paper was illusory. All the accounts we have considered are equally good, and what makes them good or correct is not their correctness definition-by-definition, as Frege would perhaps have had us believe, but properties of the completed account, *as a whole*, shared by different accounts with different definitions and which identify different sequences of objects as "the numbers." In brief, their status as *reductions* defines the standard on which they are to be judged, and on that standard they fare equally well.

This would be tantamount to concluding that there was no truth of the matter *because* the conditions on the correctness of an account expressed all the determinants of truth that there were, and these clearly were such as to determine the answer only up to isomorphism with what we know as the numbers, to put it a bit too realistically. A correct reduction is one that gives us back all that we know about the numbers and, from a structural point of view, we know all there is to know (there are certain matters of detail that still escape us, but the condition on reductions is insensitive to those differences). Hence the entreprise of reduction is precluded in advance from making finer discriminations among number-candidates than we now make among numbers themselves. This conclusion is the premiss from which Quine *began* in his earlier work on reduction and ontological commitment, and represents what I spoke of at the outset of this paper as the "instrumentalist" view of the conclusion that numbers are sets: the view that sets can "do the work of numbers" and thus are fully equipped to replace them. I will return shortly to a consideration of this early Quinean position, for it represents an alternative interpretation of the conclusion that numbers are sets, one which, on some interpretations at least, never saw the quest as one which inquired into the true nature of the numbers. But this conclusion depends upon that "instrumentalist" view of reduction and is better discussed along with its parent premiss.

Or one might think of abandoning the quest because one saw that it would never be fulfilled: that, first of all, whatever further facts we might learn about numbers *qua* numbers, we can now see that we can immediately transfer to whatever surrogates we might have chosen in a reduction. They would therefore be of no help in identifying one sequence over another as the true numbers. They would simply be further characterizations of all ω-sequences. So more number theory won't help. And secondly, if we wanted to discriminate one ω-sequence from others as the genuine numbers, we would always be hard put to answer the skeptic. The argument would go:

*We*: S is not the numbers, because Φ(S) and not-Φ(Nn)

*He*: But, how do you know that not-Φ(Nn), for if Nn = S, surely Φ(Nn), since, as you have shown, Φ(S).

If we are *now* at a standstill, we are unlikely to make much headway. The reason for this is that there seems to be now no *general* theory, discoveries within which force or even incline us to choose certain identifications over others. Of course, we might develop a theory which, did we have it now, would fit that description. But unless such a theorizing could be seen as the natural outgrowth of present theories, which don't seem to exist, it is hard to see how it could be said to answer the present question: Which sequence of objects is *really* the numbers?

Still, one might suppose that there is a glimmer of hope in this direction, offered by the fact that there do seem to be reasons to rule out *some* sequences, or at least some things, from being numbers. The reasons are a bit obscure, but, on the whole correct—I think. What I have in mind is that we can put to rest our old concern that that same familiar conqueror of Gaul might be the number 3, or any other number, for that matter. He could not. I assume that we still cling to the belief that we know that $2 < 3$. But this was known to those who had no idea of Caesar's existence—who knew nothing of Caesar. Put in this way, it seems to beg the question, for our skeptic might simply reply that if Caesar is the number 3, then of course they knew of Caesar (they knew a whole lot about Caesar) that he was divisible into three parts, that he was the square root of the Taj Mahal, etc. But now this will not do. We would like there to be an account of *how* they knew. But that would be asking too much, since we don't really have an account of how *we* know that $3 > 2$. But given that we do know about Caesar and the kind of object that he is (was) (or at least what we take him to be), we have good reason to think that the ancient Babylonians could not have known about him. And if they didn't know about him, they could hardly have known that he was >2. But they *did* know that $3 > 2$. This sounds misleadingly like a fallacious argument that depends on substituting in an opaque context. But it is not. What is in question is our conception of what we know about Caesar and how we know it. If "3" refers to Caesar, and did for the Babylonians, then it came to do so in strange and wondrous ways. To accept the possibility that they had knowledge of Caesar sufficient to yield the relevant arithmetic truths would force us to revise too much of the background theory we have about *reference* and *knowledge* in terms of which we account for our knowledge of the physical world. The obstacles, it would appear, are well nigh insurmountable—which is not to say that someone might not prefer to throw all that over for the pleasure of thinking of Julius as a number, but it is to suggest that we have enough background in our theory of knowledge and theory of reference to render the revisions they would undergo if we were to contemplate allowing Julius Caesar numerical status not worth the effort. But, as I said before, if this is any hope at all, it is only the faintest glimmer. For if these considerations have the slightest cogency, they owe it entirely to a causal account of knowledge and, underlying that, reference, which in turn is plausible only in connection with physical objects. So

using this, we cannot choose between competing *abstract* candidates for the ultimate Pythagorean reality, unless, of course, the abstractions can be sorted out on the basis of our ordinary empirical knowledge of everyday objects.

Finally, one might think of abandoning the search because one felt the question not to be a genuine one. A verificationist of sorts might conclude that it was not a real question precisely on the grounds given a page or so ago: that the answer (if any) must forever elude us. But one need not reject the question for such *prima facie* verificationist reasons. One can, like the Quine of "Ontological Relativity" be a bit more circumspect. There, he finally accepts the conclusions which have been inherent in *Word and Object* and which lead him to the doctrine of the relativity of ontology and, with it, identity, for, as he says

> [w]e cannot know what something is without knowing how it marked off from other things. Identity is of a piece with ontology. Accordingly, it is involved in the same relativity.
>
> Quine [1968] (1969: 55)

So, relativistic conclusions concerning the numbers—that they may with equal justice be identified with any sequence, lead to a relativistic interpretation of the predicate in "1 = [Λ]." What I described as a more circumspect approach to this relativistic conclusion—to this rejection of the Fregean question, realistically construed—is the grounds on which Quine rests the view that any sequence would do. In the end, it comes down to verificationism, not a general all-encompassing form of it, but a more subtle one. Quine takes his verificationist stand first in the theory of meaning, and finally in the theory of reference. So, what ontological relativity comes down to is the impossibility of concocting behavioral tests to discriminate among different analytical hypotheses all of which meet certain basic conditions. Perhaps this verificationist foundation for the theory of reference suffices to yield the more thoroughgoing variety adumbrated by the early positivists and mentioned early in this paragraph as a view contrasting with Quine's. But that would have to be shown.

These, then, are the reasons why, beginning with a realistic interpretation of the Fregean question, we would be inclined to consider whether a misconstrual of identity is not what is at fault here.

The early Quine held a view somewhat as follows. It is pieced together from a variety of sources—principally memory, so do not ask for chapter and verse.

*Ontological commitment*: this is the touchstone of the earlier views. "What is there?" There is everything a true theory says there is (and possibly more). But we can agree at least on that. So ontology reduces to: (i) forming a theory and assessing what it says there is, and (ii) determining if the theory is true. That may not be all there is to ontology—or at least it may not get a complete answer to the question (complete but more informative and detailed than "Everything"), but it is a start. So what does a theory say there is? To what is a theory committed (ontologically)? It is committed to all those things that must be counted among the values of its variables of quantification if the theory is to be true. All this would be straightforward enough, and useless enough, were it not for the word "must." For it espresses the

possibility that what may on first reckoning be needed as a value of a bound variable might actually, on deeper probing, prove superfluous. Hence the important notion of paraphrase. If a sentence (theory) makes *apparent* reference to *F*'s, but can be paraphrased into one that does not need to count any *F*'s among the values of its variables to be reckoned as true, then the original reference was apparent only.

Witness:

(1) 5 is the number of planets.

That would seem to imply that (∃x) (x is a number). Thus, our original would appear to be committed to numbers. But we all know that such is not the case. For (1) can be paraphrased into a sentence whose quantifiers range only over bona fide physical objects—planets—and not numbers at all (unless *they* are planets—which raises the question of the opacity or transparency of the criterion).

You will immediately notice two things about this criterion: (i) it depends on some notion of paraphrase, and (ii) it contains a built-in reductionist feature. Some form of Occam's razor is guiding the direction of the reduction. The ontology of a theory is the *sparest* it can do with. It is the original apparent ontology stripped by paraphrase of all inessentials. Try as one may to be lavish in one's ontology, one cannot. For the criterion dictates that if our apparent commitments can be pared away, they are apparent only. The criterion stands as our ontic super-ego. Although at the outset it may have seemed neutral, it ends up with a strong reductionist bias. Applied to the case most immediately at hand, it is unclear whether the reduction of arithmetic to set theory shows that a theory which contains number-expressions as well as set-expressions (but hasn't defined the one in terms of the other) is uncommitted to numbers, or that it *is* committed to numbers, but through its set vocabulary. For to what are you committed when you *are* committed to numbers, say, in number theory? Well, to things satisfying the axioms. But what is it to satisfy the axioms? Does only the intended model satisfy the axioms? Or do other things do so as well? Clearly, the latter is intended. For if the former were intended, no reduction or ontology would be possible. Hence the ontology of a pure number theory is *itself* not specifiable except in some stronger theory which tells us more about the range of the variables and the extensions of the predicates. But now we are closing in on the position of "Ontological Relativity," where what number theory is said to be committed to is, quite evidently, relative to the theory in which the reduction is carried out.

What is at stake here is the notion of paraphrase and its neutrality. In the earlier papers, Quine's view had a realistic ring. A theory has certain commitments, and it is either true or false. In *Word and Object*, the notion of translation into canonical notation is developed as the successor or *explicans* of paraphrase, and it is explicitly recognized that different translations might be carried out for different purposes, and hence that there may be no fact of the matter concerning the ontology of a theory. In "Ontological Relativity," these views are pushed to their extremes by undercutting the concept of truth entirely.

So reduction, seen as casting about for what would do the work of the numbers, also leads to a relativistic view of the class of identities Frege wanted to settle. For the identification of the "job to be done" is either made in terms of the truth theory for the language—*the* theory or *a* theory that attributes an ontology to it—in which case there is no reduction to be had and only the numbers can do their job, or else it is done in some way that eschews outright identification of the objects and settles for isomorphisms to define the job, in which case there is no pretense of identifying the objects through finding what would do their job.

**Note**

1. See Benacerraf (1965). *Editor's note.*

## Reference

Quine, W. V. O. [1968] (1969). Ontological relativity. In *Ontological relativity and other essays* (pp. 26–68). New York: Columbia UP.

# Chronological Bibliography of Paul Benacerraf to 2016 (to the exclusion of reprints in anthologies)

## Dissertation

(1960). *Logicism, Some Considerations*. Ph.D. Dissertation, Princeton University, 260 pages + Abstract; Preface: pp. 1–2, Introduction: pp. i–vi, Appendix: pp. A-1–A-15 [Reference: 61–4511]. Ann Arbor: University Microfilms Inc.

## Articles

(1962). Tasks, Super-Tasks, and Modern Eleatics. *The Journal of Philosophy, 59*(24), 765–784.
(1963). Mr. Searle on Meaning and Speech Acts. In C. D. Rollins (Ed.), *Knowledge and Experience*, *Proceedings of the 1962 Oberlin Colloquium in Philosophy* (pp. 43–49), with a rejoinder by John R. Searle at pp. 50-54). Pittsburgh: University of Pittsburgh Press.
(1965). What Numbers Could Not Be. *The Philosophical Review, 74*(1), 47–73.
(1967). God, the Devil and Gödel. *The Monist, 51*, 9–32.
(1973). Mathematical Truth. *The Journal of Philosophy, 70*(19), 661–679.
(1981). Frege: The Last Logicist. In P. A. French, T. E. Uehling Jr. and H. K. Wettstein (Eds.), *The Foundations of Analytic Philosophy*, *Midwest Studies in Philosophy, 6*(1), 17–35.
(1984). Comments on Maddy and Tymoczko. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 2*, 476–485.
(1985). Skolem and the Skeptic. *Proceedings of the Aristotelian Society,* Supplementary vol. *59,* 85–115.
(1994). Paul Ziff, 1958–60: A Reminiscence. In D. Jamieson (Ed.), *Language, Mind and Art: Essays in Appreciation and Analysis in Honor of Paul Ziff* (pp. 1–7). The Netherlands: Kluwer Academic Publishing, Synthese Library.
(1996a). Disputation. *Philosophia Mathematica, 4*(2) [S. Shapiro (Guest Ed.), LOGICA 95 - *Proceedings of the 5th Congress of the Academy of Sciences of the Czech Republic*, June 1995], 184-189; reprinted as "Appendix: Recantation *or* Any old ω-sequence would do after all" to *Benacerraf 1996b*, at pages 45–48.
(1996b). What Mathematical Truth Could Not Be—I. In A. Morton and S. P. Stich (Eds.), *Benacerraf and his Critics* (pp. 9–59). Oxford, and Cambridge, Mass.: Blackwell Publishers.
(1999). What Mathematical Truth Could Not Be—II. In S. Barry Cooper and J. K. Truss (Eds.), *Sets and Proofs,* London Mathematical Society Lecture Note Series *258* (pp. 27–51). Cambridge: Cambridge UP.



F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology,
and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6

(2016a). Mathematical Truth (January 1968 version). In F. Pataut (Ed.), *Truth, Objects, Infinity: New Perspectives on the Philosophy of Paul Benacerraf* (pp. 263–287, in this volume). Berlin: Springer, Logic, Epistemology and the Unity of Science Series.
(2016b). Comments on Reduction (Lecture in a graduate seminar ∼1975). In F. Pataut (Ed.), *Truth, Objects, Infinity: New Perspectives on the Philosophy of Paul Benacerraf* (pp. 289–296, in this volume). Berlin: Springer, Logic, Epistemology and the Unity of Science Series.

## Editor

[1964] (1983). P. Benacerraf and H. Putnam (Eds.), *Philosophy of mathematics—Selected readings*. 1st ed., Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1964; 2nd ed., revised with a new preface, a new introduction and a new bibliography, Cambridge, Mass.: Cambridge UP, 1983.

## Translations

(1980a). *Essay on the Causal Theory of Time* (English translation, revised by C. R. Fawcett and R. S. Cohen, of Henry Mehlberg's *Essai sur la théorie causale du temps*). Vol. 1 of R. S. Cohen (Ed.), *Time, Causality, and the Quantum Theory*, with a preface by A. Grünbaum. Dordrecht: D. Reidel Publishing Co.
(1980b). With R. C. Jeffrey. On the condition of partial exchangeability. (English translation of "Sur la condition d'équivalence partielle" by Bruno de Finetti.) In R. Carnap and R. C. Jeffrey (Eds.), *Studies in Inductive logic and Probability* (Vol. II, pp. 193–205). Los Angeles: University of California Press.

## Articles, Preface and Introduction in Collaboration

[1964] (1983). With Hilary Putnam. Introduction to *Philosophy of mathematics—Selected readings* [1st ed., Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1964]; Preface to the second edition, Introduction, and a new revised bibliography, originally compiled by George Boolos and augmented by Tagashi Yagasawa. Cambridge, Massachusetts: Cambridge UP, 1983 [Preface: pp. vii–viii, Introduction: pp. 1–37, Bibliography: pp. 571–600].
(1977). With Simon Kochen and Gerald Sacks. Meeting of the Association for Symbolic Logic: New York, 1975. *The Journal of Symbolic Logic, 42*(1), 143–155.
(1998). With Richard Jeffrey. Carl Gustav Hempel 1905–1997. *Proceedings and Addresses of the American Philosophical Association, 71*(5), 147–149.

## Miscellaneous

(1964). Review: Paul Ziff, Semantic Analysis. *The Journal of Symbolic Logic, 29*(4), 193–194.
(1972). With W. G. Bowen, T. A. Davis, W. W. Lewis, L. K. Morse, and C. W. Schafer. *Budgeting and Resource Allocation at Princeton University.* Report of a Demonstration Project Supported by the Ford Foundation, Princeton, New Jersey, June [390 pages + Appendices].
(1998). Margaret Dauler Wilson 1939–1998. *Proceedings and Addresses of the American Philosophical Association, 72*(2), 126–127.

(2000). An Interview with Paul Benacerraf [Questions by Mark Crimmings, Grigory Mints, Richard Zach and Edward Zalta from the Philosophy Department of Stanford University, and by the editors of *The Dualist*]. *The Dualist*, June, 59–65.

## Translated Articles and Translations of Benacerraf and Putnam (Eds.), [1964] (1983)

(1976). Mathematische Warheit. (German translation of "Mathematical Truth" by H. Vetter.) In M. Sukale (Ed.), *Moderne Sprachphilosophie* (pp. 216–232). Hamburg: Hoffman und Campe Verlag.

(1993a). Ἡ μαθηματικὴ ἀλήθεια. (Greek translation of "Mathematical Truth".) In P. Christodoulidis (Ed.), *Η ΦΙΛΟΣΟΦΙΑ ΤΩΝ ΜΑΘΗΜΑΤΙΚΩΝ [The Philosophy of Mathematics]* (pp. 241–261), published under the auspices of the school of philosophy of the university of Ioannina, Athens.

(1993b). Qué no Podrian ser los Numeros. (Spanish translation of "What Numbers Could not Be".) *Mathesis, 9*(3), 317–343.

(1995). Suugaku-teki Shinri. (Japanese translation of "Mathematical Truth" by T. Iida.) In T. Iida (Ed.), *Riindingusu Suugaku no Tetsugaku: Geederu Igo* [*Readings in the Philosophy of Mathematics: After Gödel*] (pp. 245–272). Tokyo: Keiso Shobo.

(2001). *Hyondae sahoe eso kongdongche nun kanung hanga (Taeu haksul chongso).* (Korean translation and edition of *Philosophy of mathematics—Selected readings*), Tae-gi Kand (Ed.). Seoul: Akanet.

(2003a). *Shu xue zhe xue*. (Chinese translation and edition of *Philosophy of mathematics—Selected readings*), Zhu Shui Lin (Ed.). Beijing: Shang wu yin shu guan Publisher.

(2003b). Amik A Szamok Nem Lehetnek. (Hungarian translation of "What Numbers Could Not Be" by G. Farkas.) In F. Csaba (Ed.), *A matematiks filozofiaja* (pp. 1–34). Budapest: Osiris.

(2004). La Verdad Matemática. (Spanish translation of "Mathematical Truth" by P.-B. Fornés Ferrer & F. Santonja Gómez, revised by F. Rodríguez-Consuegra, with an introductory note by F. Rodríguez-Consuegra.) *Ágora - Papeles de Filosofia, 23*(2), 233–253.

(2014a). Ce que les nombres ne peuvent pas être. (French translation of "What Numbers Could Not Be" by S. Maronne.) In *Philosophie des mathématiques - Ontologie, vérité et fondements*, Textes réunis par S. Gandon et I. Smadja (pp. 45–73). Paris: Librairie Philosophique Jean Vrin, coll. "Textes clés".

(2014b). La vérité mathématique. (French translation of "Mathematical Truth" par B. Halimi.) In *Philosophie des mathématiques - Ontologie, vérité et fondements*, Textes réunis par S. Gandon et I. Smadja (pp. 75–96). Paris: Librairie Philosophique Jean Vrin, coll. "Textes clés".

## Unpublished Manuscripts

(2000). Should We Believe 'Must We Believe in Set Theory?'. The paper is an improved version of "What Mathematical Truth Could Not Be - II" ((1999) in the Articles section of this Chronological Bibliography). Various amended versions of the 1999 article have been presented at UCLA, Pomona College, Columbia University and the University of Athens. The last version, dated 2000, is slightly improved from the lecture delivered in Paris at the Séminaire de Philosophie des Mathématiques et de l'Informatique of the Institut d'Histoire et de Philosophie des Sciences et des Techniques on October 11, 1999. Boolos's article, "Must We Believe in Set Theory?", was originally published in George Boolos: *Logic, Logic and Logic*, With Introductions and Afterword by John P. Burgess, Edited by Richard Jeffrey, Harvard UP, Cambridge, Mass., 1998, pp. 120–132. [33 pages]

(2016). Hilary Putnam. Obituary read on March, 14, 2016 at Levine Chapel, Brookline, Mass. on the occasion of Hilary Putnam's funeral. [7 pages]

# Author and Citation Index





F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6

# Name Index





F. Pataut (ed.), *Truth, Objects, Infinity*, Logic, Epistemology, and the Unity of Science 28, DOI 10.1007/978-3-319-45980-6